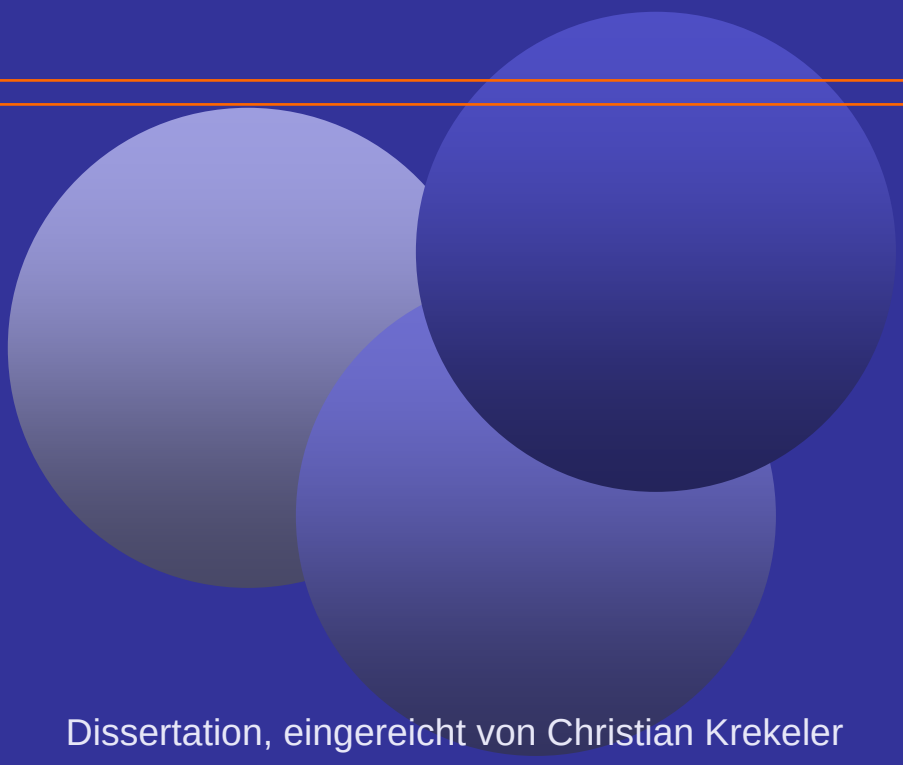
Grammatik und Fachbezug in Sprachtests für den Hochschulzugang



Dissertation, eingereicht von Christian Krekeler
beim Fachbereich für Geisteswissenschaften
der Universität Duisburg-Essen
zur Erlangung der Würde eines Dr. phil.

Dezember 2005

Grammatik und Fachbezug in Sprachtests für den Hochschulzugang

Dissertation,
eingereicht von Christian Krekeler (geboren in Rosendahl, NRW)
beim Fachbereich für Geisteswissenschaften
der Universität Duisburg-Essen
zur Erlangung der Würde eines Dr. phil.

Betreuung der Arbeit und Erstgutachter: Prof. Dr. Rupprecht S. Baur
Zweitgutachter: Prof. Dr. Peter Raster
Vorsitzende des Promotionsausschusses: Prof. Dr. Emel Huber
Datum der Disputation: 14. Dezember 2005

Inhaltsverzeichnis

Verzeichnis der Abbildungen

Verzeichnis der Tabellen

Abkürzungen

AM	arithmetisches Mittel
CPE	*Certificate of Proficiency in English*
DSH	Deutsche Sprachprüfung für den Hochschulzugang
EAP	*English for Academic Purposes*
EEP	*English for Educational Purposes*
EGAP	*English for General Academic Purposes*
ELTS	*English Language Testing System (Test)*
EOP	*English for Occupational Purposes*
ESAP	*English for Special Academic Purposes*
EST	*English for Science and Technology*
ETS	*Educational Testing Service*
FaDaF	Fachverband Deutsch als Fremdsprache
FCE	*First Certificate in English*
HV	Hörverstehen
IELTS	*International English Language Testing System (Test)*
LV	Leseverstehen
MA	Mündlicher Ausdruck
Md	Median
MLAT	*Modern Language Aptitude Test*
Mo	Modalwert
MP	Mündliche Prüfung
N, n	Anzahl (der gesamten Stichprobe bzw. einer Teilgruppe)
p	Signifikanz
PNdS	Prüfung zum Nachweis deutscher Sprachkenntnisse
$\mathbf{r^2_{KORR}}$	korrigiertes R-Quadrat
$\mathbf{r_s}$	Rangkorrelationskoeffizient nach Spearman
s	Standardabweichung
SA	Schriftlicher Ausdruck
SPEAK	*Speaking Proficiency English Assessment Kit*
TDN	TestDaF-Niveaustufe
TEACH	*Taped Evaluation of Assistants' Classroom Handling*
TestDaF	Test Deutsch als Fremdsprache für ausländische Studienbewerber
T-Kurs	Kurs an Studienkollegs für Studienbewerber, die ein technisches Studium anstreben

TOEFL	*Test of English as a Foreign Language*
TP	Textproduktion
TSE	*Test of Spoken English*
UCLES	*University of Cambridge Local Examination Syndicate*
W-Kurs	Kurs an Studienkollegs für Studienbewerber, die ein wirtschaftswissenschaftliches Studium anstreben

Dank

Mein Dank gilt den ausländischen Studienbewerberinnen und Studienbewerbern für ihre Bereitschaft, an den Tests und Umfragen teilzunehmen, und den Kolleginnen und Kollegen an der Fachhochschule Konstanz und an mehreren Studienkollegs, welche mich bei der Durchführung der Erhebungen unterstützten.

Prof. Dr. Rupprecht S. Baur danke ich für die Betreuung der Arbeit und die langjährige Unterstützung.

1. Einleitung

Die Werbung ist erfolgreich, das Produkt nicht. Die Anwerbung ausländischer Studierender im Ausland durch den DAAD und deutsche Hochschulen zeigt Früchte, der Studienerfolg ausländischer Studierender an deutschen Hochschulen ist jedoch mäßig (DAAD, 2000; 2003). Die Zahl der ausländischen Studierenden in Deutschland stieg von 1997 bis 2002 um über 40 Prozent. Noch deutlicher wird der Trend, wenn man die Zahl der ausländischen Studien*anfänger* an deutschen Hochschulen betrachtet. Zwischen 1997 und 2001 war eine Zunahme von über 70 Prozent zu verzeichnen (siehe Tabelle 1 und Abbildung 1). Dies ist auch ein Ergebnis der zunehmenden Internationalisierung der Bildungslandschaft.

Laut einer vom DAAD in Auftrag gegebenen Untersuchung machen aber nur ungefähr 30 Prozent der ausländischen Studienanfänger eine Abschlussprüfung an der Hochschule, an der sie das Studium aufnahmen. Unter Berücksichtigung der Studienwechsler dürfte – so wird in der Studie spekuliert – ungefähr die Hälfte der ausländischen Studierenden in Deutschland auch einen Studienabschluss erreichen (DAAD, 2003; Heublein/Sommer/Weitz, 2004). Angesichts der steigenden Zahlen und des geringen Studienerfolgs kommt der sprachlichen Vorbereitung auf ein Studium an einer deutschen Hochschule eine große Bedeutung zu. Dass die Sprachkompetenz ausländischer Studierender ein Schlüssel für den Studienerfolg darstellt, ist unbestritten (z. B. Blue, 1993); legitim und notwendig ist daher, dass sich die aufnehmenden Hochschulen vor der Aufnahme ausländischer Studienbewerber ein Bild über deren Deutschkompetenz machen.

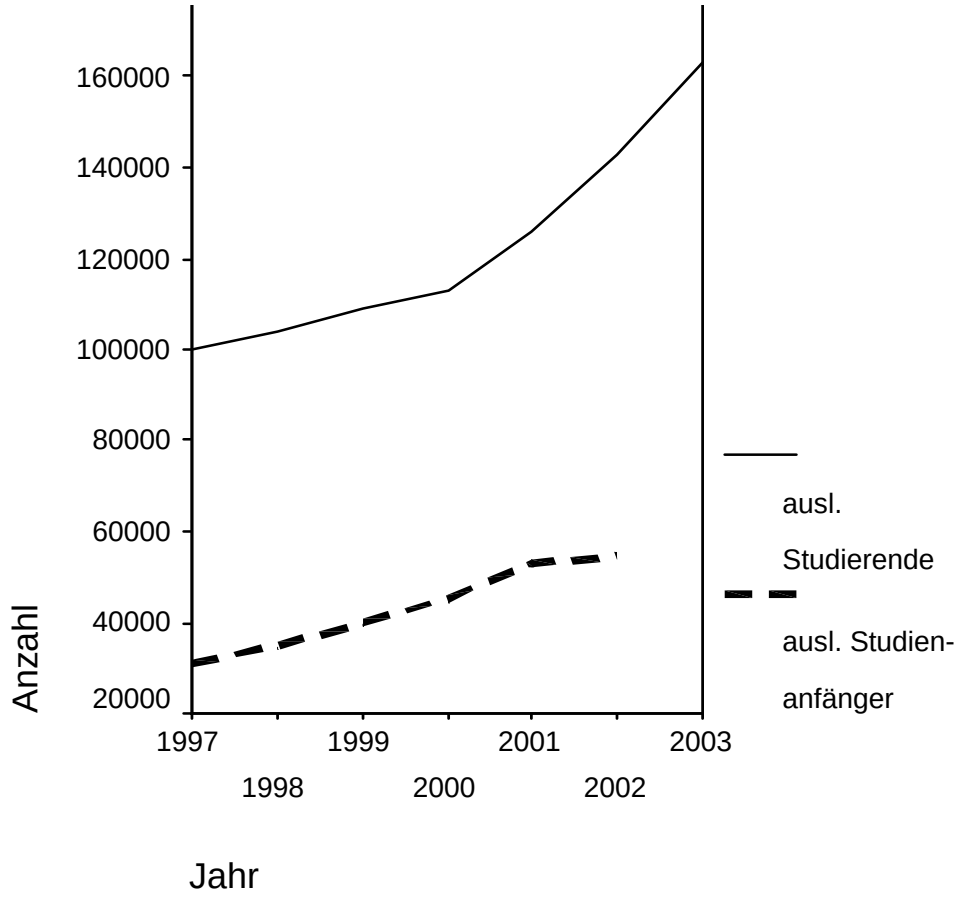
In Deutschland gibt es zwei Prüfungen, welche für den Nachweis ausreichender Deutschkenntnisse für ein Fachstudium konzipiert wurden: die "Deutsche Sprach-

prüfung für den Hochschulzugang ausländischer Studienbewerber" (DSH) und den "Test Deutsch als Fremdsprache für Studienbewerber" (TestDaF).

Tabelle 1: Ausländische Studierende an deutschen Hochschulen (nur "Bildungsausländer")

Jahr	1997	1998	1999	2000	2001	2002	2003
ausländische Studierende an deutschen Hochschulen (ohne "Bildungsinländer")	100.033	103.716	108.785	112.883	125.714	142.786	163.213
ausländische Studienanfänger an deutschen Hochschulen (ohne "Bildungsinländer")	31.125	34.760	39.898	45.149	53.157	54.480	

Quelle: DAAD/HIS, 2004; HIS, 2005.



Quelle: DAAD/HIS, 2004; HIS, 2005.

Abbildung 1: Ausländische Studierende und ausländische Studienanfänger an deutschen Hochschulen (ohne "Bildungsinländer") – Liniendiagramm

Alternative Prüfungen wie die "Zentrale Oberstufenprüfung" (ZOP), das "Kleine Deutsche Sprachdiplom" (KDS), das "Große Deutsche Sprachdiplom" (GDS) sowie das "Deutsche Sprachdiplom der Kultusministerkonferenz, Stufe II" (DSD II) werden weiterhin als Studienvoraussetzung anerkannt. Im Gegensatz zur DSH und zum TestDaF beziehen sich diese Tests nur am Rande auf studienspezifische Sprachverwendungssituationen. Die Anerkennung als Nachweis ausreichender Sprachkenntnisse für ein Studium in Deutschland ist ein Nebeneffekt dieser Tests, nicht ihr Hauptanliegen. Mit der Überarbeitung der Richtlinien für deutsche Sprachprüfungen (HRK/KMK, 2004) wurde deutlich gemacht, dass der TestDaF und die DSH gleichberechtigt als Tests zum Nachweis von Sprachkenntnissen für den Hochschulzugang angesehen werden. Die DSH wird jährlich etwa 15.000 bis 20.000 Mal abgenommen (Casper-Hehne/Koreik, 2002). Die Zahl der Teilnehmer am TestDaF ist derzeit noch geringer, sie nimmt aber rapide zu: von 1.000 Teilnehmern im Jahre 2001 über 3.200 im Jahre 2002, 7.500 im Jahre 2003 auf 8.900 im Jahre 2004 (Auskunft des TestDaF-Instituts).

Wer die sprachwissenschaftliche Diskussion um Sprachtests verfolgt, trifft im englischen Sprachraum auf umfangreiche Forschungen, auch im deutschsprachigen Raum gibt es eine Vielzahl von Publikationen zu Sprachtests (Forschungsübersichten in: ALTE, 2001; Bachman, 2000; Grotjahn, 2004; Kunnan, 1999). Relativ unberührt von Begleitforschung war bis vor einigen Jahren die DSH, die bislang größte deutsche Sprachprüfung für ausländische Studienbewerber. Es wurden nur wenige Dissertationen erstellt oder Beiträge in Fachzeitschriften veröffentlicht, die sich mit dem deutschen Sprachtest für den Hochschulzugang beschäftigen (Ausnahme: Lee, 1998). Ein Grund für das Desinteresse liegt möglicherweise in der fehlenden Standardisierung der DSH: Untersuchungen zu *einer* DSH treffen mit großer Wahrscheinlichkeit auf viele andere nicht zu. Das relative Desinteresse der sprachwissenschaftlichen Forschung an Sprachprüfungen für den Hochschulzugang im deutschsprachigen Raum gehört spätestens seit der Entwicklung des TestDaF und der Gründung des TestDaF-Instituts der Vergangenheit an. Zum relativ neuen TestDaF existiert eine bereits ansehnliche Begleitforschung (siehe www.testdaf.de). Die vorliegende Arbeit ist auch eine Folge der neuen Aufmerksamkeit, die deutschen Sprachtests gegenwärtig gewidmet wird. Eine Folge der neuen Aufmerksamkeit für Sprachtests ist offensichtlich auch, dass sie häufig überarbeitet werden. Während diese Arbeit entstanden ist, wurde die DSH einigen Veränderungen unterzogen. Bei der Endredaktion im März 2005 wurde die DSH-Rahmenordnung von

2004 berücksichtigt. Das DSH-Handbuch zur neuen Rahmenordnung, das für Anfang 2006 angekündigt ist, jedoch noch nicht.

Tests und Prüfungen rufen häufig negative Gefühle hervor, denn durch die Selektionsfunktion wird mit dem Einsatz von Prüfungen auch Macht ausgeübt. Auch als Reaktion auf die Bedenken gegenüber den negativen Eigenschaften von Tests und Prüfungen wurden alternative Formen entwickelt, mit denen man Sprachkenntnisse erfassen kann (Perlmann-Balme, 2001; Brown/Hudson, 2002). Die Selbstvergewisserung über ein Portfolio oder die Selbsteinschätzung sind zwei Beispiele dafür (Black, 1994; Brown, 1998). Derartige Verfahren gehören jedoch nicht zum Thema der vorliegenden Arbeit, die sich mit einem eher klassischen Thema des Testens von Sprache befasst: Sprachtests mit gewichtigen Konsequenzen für die Teilnehmer.

Ausgangspunkt dieser Arbeit sind Besonderheiten der DSH und Perspektiven für deutsche Sprachtests für den Hochschulzugang. Zwei Aspekte, die in standardisierten Sprachtests für den Hochschulzugang an Bedeutung verloren haben, sind Grammatik und Fachbezug. Dieser Trend manifestiert sich am neu entwickelten TestDaF: Ein Grammatiktest ist nicht enthalten, auf einen Fachbezug verzichtet man. Bei der DSH werden beide Aspekte ebenfalls diskutiert: Der Umfang des DSH-Grammatiktests wurde mit der neuen DSH-Rahmenordnung von 2004 reduziert; ein Fachbezug ist nur unter bestimmten Umständen vorgesehen (HRK/KMK, 2004). Im Rahmen dieser Arbeit interessiert mich, ob und unter welchen Bedingungen diese beiden Aspekte eine Berechtigung in Sprachtests für den Hochschulzugang haben und die Nützlichkeit der Prüfungen erhöhen könnten. Möglicherweise stellen Grammatik und Fachbezug eine Chance für die Weiterentwicklung der DSH als zielgruppengenaue Prüfung mit hoher Authentizität und Praxisnähe dar, möglicherweise werden diese Aspekte im TestDaF nicht ausreichend gewürdigt.

Aufbau der Arbeit: Bevor ich mich mit den beiden Schwerpunktthemen der Arbeit beschäftige, mache ich eine "Bestandsaufnahme", in der beide deutsche Sprachtests für den Hochschulzugang vor dem Hintergrund testtheoretischer Konzepte vorgestellt werden (Kapitel 2). Eine Vertrautheit mit den Testformaten setze ich dabei voraus. Die allgemeine Bestandsaufnahme bereitet den Hintergrund für die folgenden Kapitel, in denen Detailfragen zu Sprachtests mit Hochschulzugang erläutert und erforscht werden. Die Schwerpunktthemen dieser Arbeit, Grammatik und Fachbezug in Sprachtests für

den Hochschulzugang, ergeben sich aus Unterschieden zwischen der DSH und dem TestDaF.

Während viele große Sprachtests auf einen expliziten Test der grammatischen Kenntnisse verzichten, enthält die DSH einen expliziten Grammatiktest. Dieser Sachverhalt verleitete mich dazu, Fragen nach der Nützlichkeit eines Grammatiktests zu stellen. Der DSH-Grammatiktest soll auch nach einer Überarbeitung beibehalten werden, allerdings mit einem verringerten Umfang und als Teil eines anderen Prüfungsteils, dem Leseverstehen (HRK/KMK, 2004). Diese Änderungen deuten auf ein gewisses Unbehagen über den Grammatiktest: Man möchte nicht ganz darauf verzichten, ihn jedoch auch nicht zu stark bewerten. Diese Unsicherheiten im Umgang mit dem Grammatiktest sind nachvollziehbar. Im Falle von IELTS, einem englischen Sprachtest für den Hochschulzugang, wurde der Grammatiktest gestrichen, nachdem man ihn auf seine Aussagekraft untersucht hatte (Alderson, 1993). Auch beim TOEFL wird man in Zukunft auf den Prüfungsteil *Structure and Written Expression* verzichten. Vor diesem Hintergrund frage ich, welchen Nutzen, welche Auswirkungen ein Grammatiktest in einem Sprachtest für den Hochschulzugang hat. Ist ein Grammatiktest überhaupt zu legitimieren? Andererseits ist zu fragen, welche Auswirkungen der Verzicht auf einen Grammatiktest im TestDaF hat. Diesen Themen gehe ich in den Kapiteln 3 und 4 nach.

Auch beim Thema "Fachbezug in Sprachtests für den Hochschulzugang" war der Unterschied zwischen der DSH und dem TestDaF Ausschlag gebend für mein Interesse. Die DSH-Rahmenordnung räumt DSH-Ausrichtern die Möglichkeit ein, unter bestimmten Umständen einen Fachbezug herzustellen (FaDaF, 2001: 3/4). Diese Möglichkeit scheint nur in Ausnahmefällen wahrgenommen zu werden, regelmäßig wird ein Fachbezug dagegen im Deutschteil der Feststellungsprüfung hergestellt. Ist das ein Modell für Sprachprüfungen für den Hochschulzugang? Beim TestDaF verzichtet man grundsätzlich auf einen Fachbezug. Bei IELTS führte die umfangreiche Begleitforschung dazu, dass man ebenfalls auf einen Fachbezug verzichtete (Clapham, 1996; 2000). Es gibt aber auch eine entgegen gesetzte Entwicklung: Es wurden seit den 1990er Jahren mehrere Sprachtests mit Fachbezug konzipiert, dabei handelt es sich jedoch vor allem um Sprachtests für berufliche Zwecke. Das Forschungsinteresse an Sprachtests mit Fachbezug ist groß, wie z. B. aus der Monographie von Douglas hervorgeht (Douglas, 2000). Die Diskussion um Sprachtests mit Fachbezug fasse ich in Kapitel 5 zusammen,

über eine von mir zu diesem Thema durchgeführte Studie berichte ich in Kapitel 6. In dieser Studie wird eine Frage untersucht, die für die Validität von Sprachtests mit Fachbezug von zentraler Bedeutung ist: Unter welchen Bedingungen können Testteilnehmer ihre Vorkenntnisse zum Fachthema einbringen? Hängt es, so frage ich genauer, vom Niveau der Fremdsprachenkenntnisse ab, ob etwaige Vorkenntnisse eingesetzt werden können? Wenn sich der Einfluss der Vorkenntnisse etwa in Abhängigkeit vom Niveau der Fremdsprachenkenntnisse beschreiben ließe, könnte dies bei der Testerstellung berücksichtigt werden.

In der vorliegenden Arbeit stehen zwei Aspekte im Mittelpunkt: der Umgang mit Grammatik in Sprachtests für den Hochschulzugang sowie ein möglicher Fachbezug. Dabei traf ich auf eine Reihe von Querverbindungen:

- Bei beiden Themen handelt es sich um Unterschiede zwischen der DSH und dem TestDaF.

- Beide Themenkreise – Grammatik und Fachbezug in Sprachtests für den Hochschulzugang – wurden mit Blick auf einen englischen Sprachtest für den Hochschulzugang, IELTS, in umfangreichen Studien diskutiert (siehe Kapitel 3).

- Ein weiterer Zusammenhang zwischen beiden Themen wird in einer Argumentation von Clapham (2000: 519) deutlich: Sie fordert die Entwicklung von Eignungstests anstelle der derzeit eingesetzten Feststellungsprüfungen. Derweil sollte man ihrer Ansicht nach statt Sprachtests mit Fachbezug Sprachtests mit allgemeinem akademischen Inhalt zusammen mit Grammatiktests einsetzen. Ich vertrete demgegenüber den Standpunkt, dass der Einsatz von Sprachtests mit Fachbezug und der Verzicht bzw. die Reduktion von Grammatiktests eher zu einer sinnvollen Studienvorbereitung führen dürften.

Zum Sprachgebrauch: Wer Forschungsergebnisse zu Sprachtests rezipiert, muss auf englischsprachige Literatur zurückgreifen, denn viele Forschungsberichte zu Sprachtests sind auf Englisch verfasst. Wer deutsche Texte zum Testen von Sprache schreibt, muss sich entscheiden, ob die englischen Termini übernommen werden sollen oder ob eine vollständige Übersetzung der Termini erfolgen soll. Ich verwende deutsche Fachbegriffe und verweise bei erstmaliger Verwendung auf die – häufig bekannteren – englischen Pendants, damit die Orientierung erleichtert wird. Dabei beziehe ich mich auf das

Multilingual glossary of language testing terms (1998), welches als Reaktion auf die zunehmende Differenzierung der Diskussion um Sprachtests erstellt wurde. Auf Übersetzungen verzichtete ich bei Zitaten, Titeln oder fest stehenden Bezeichnungen – wie im vorhergehenden Satz. Deutschsprachige Zitate wurden gegebenenfalls in der alten Rechtschreibung belassen.

Gibt es einen inhaltlichen Unterschied zwischen einem Sprach*test* und einer Sprach*prüfung*? Der allgemeinsprachliche Umgang ist nicht trennscharf, wie aus den Erklärungen des "Großwörterbuchs Deutsch als Fremdsprache" hervorgeht:

> **Test** *der; -s, -s/-e*; **1** die Überprüfung und Bewertung bestimmter Leistungen e-r Person <ein psychologischer T.; j-n e-m T. unterziehen; e-n T. bestehen> ...
>
> **Prüfung** *die; -, -en*; **1** e-e mündliche od. schriftliche Aufgabe, mit der j-s Kenntnisse od. Fähigkeiten beurteilt werden ≈ Test, Examen <e-e mündliche, schriftliche, schwierige P.; sich auf e-e P. vorbereiten; auf / für e-e P. lernen; e-e P. machen, ablegen, schreiben, bestehen; in e-r P. versagen; durch e-e P. fallen> ... (Langenscheidts Großwörterbuch Deutsch als Fremdsprache, 1998: 978 und 775)

"Test" und "Prüfung" werden weitgehend synonym verwendet. Wenn Unterschiede vorgenommen werden, dann bezeichnet "Test" weniger formalisierte Testverfahren, welche vor Ort konzipiert und durchgeführt werden, "Prüfungen" sind dahingegen eher formalisiert und standardisiert (Perlmann-Balme, 2001). Auf diese Unterscheidung trifft man auch beim fachlichen Sprachgebrauch, dort unterscheidet man jedoch noch weitere Bedeutungen von "Test". Nach dem *Multilingual glossary of language testing terms* versteht man unter einem Test die "Prozedur zur Feststellung der fremdsprachlichen Leistungsfähigkeit" (1998: 127). Es werden zwei weitere Bedeutungsvarianten angegeben: Test als Bezeichnung für den Teil einer Prüfung (z. B. Test zum Sprechen) und Test als informelles Prüfungsverfahren. Der Bedeutungsumfang von "Prüfung" ist enger, er deckt sich mit der ersten Variante von "Test":

> Prozedur zur Feststellung der Leistungsfähigkeit oder des Kenntnisstandes von Personen durch mündliche und/oder schriftliche Aufgaben. Das Erreichen einer Qualifikation (z. B. durch ein Abschlusszeugnis oder ein Zertifikat bestätigt) oder der Zugang zu einer Ausbildung oder einem Studium etc. kann von dem Ergebnis abhängen (Multilingual glossary of language testing terms, 1998: 119).

Auch aus der testtheoretischen Definition von Lienert und Raatz wird deutlich, dass der Begriff "Test" als Fachbegriff mit vielfältigen Bedeutungen verwendet wird. Sie zählen mehrere Bedeutungen auf (z. B. auch die kurze, außerplanmäßige "Zettelarbeit" im Schulunterricht) und definieren:

> Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung (Lienert/Raatz, 1994: 1).

Ich orientiere meinen Sprachgebrauch an den Definitionen aus dem *Multilingual glossary of language testing terms*, wobei ich die Begriffe "Sprachtest" und "Sprachprüfung" synonym verwende, wenn die (umfangreiche) "Prozedur zur Feststellung der fremdsprachlichen Leistungsfähigkeit" gemeint ist. Hier gibt es keinen Anlass für eine inhaltliche Unterscheidung, was sich auch in der Bezeichnung der Sprachtests äußert: Beim TestDaF wählte man "Test", bei der ZMP "Prüfung". Für das informelle Prüfungsverfahren verwende ich "Test", für den Teil einer Prüfung "Prüfungsteil".

In der Testmethodik bezeichnet man jedes Einzelelement eines Tests, das getrennt bewertet wird, als "Item". Eine "Aufgabe" ist demgegenüber die Kombination aus Anweisung, Aufgabenstellung und Beantwortung. In einer engen Bedeutung wird häufig "Aufgabe" auch für "Item" verwendet (Multilingual glossary of language testing terms, 1998; Grotjahn, 2000a).

"Fremdsprache" oder "Zweitsprache"? Ausländische Studienbewerber sind normalerweise Lerner des Deutschen als Fremdsprache. Mit der Einreise nach Deutschland bzw. mit der Aufnahme des Studiums in Deutschland nimmt Deutsch durchaus den Status einer Zweitsprache an, da es zur Sprache der unmittelbaren Umgebung und des Unterrichts wird (Baur, 2001). Im Folgenden verwende ich "Fremdsprache" als Oberbegriff und "Zweitsprache" für den Spezialfall.

Was ist ein "C-Test"? Dieses Testformat gehört zwar nur am Rande zum Thema der Arbeit, ich beziehe mich jedoch immer wieder darauf. Ein C-Test besteht aus mehreren kurzen Texten, in denen in jedem zweiten Wort die zweite Hälfte von den Testteilnehmern rekonstruiert werden muss (Raatz/Klein-Braley, 1992: 75-77. Weitere Hinweise: Kapitel 6.1.3, Seite 255). Ein Beispiel:

Bedeutung der Lesefähigkeit
Die Lesefähigkeit trägt ihren Wert natürlich in sich, hat aber auch ökonomische Auswirkungen. Erwac_____ Leser, d_____ besser le_____ als d_____ Durchschnitt, üb_____ mit größ_____ Wahrscheinlichkeit gutbe_____ Berufe a_____. Die wach_____ Spezialisierung i_____ der Gesell_____ erfordert me_____ Bildung, ei_____ Forderung, d_____ vor al_____ an d_____ Schulen geri_____ wird. Du_____ die erhö_____ Anforderungen a_____ das Bildungsniveau, die heute in den westlichen Gesellschaften gestellt werden, ist die Lesefähigkeit des Einzelnen immer wichtiger geworden.

2. Sprachtests für den Hochschulzugang: Bestandsaufnahme

Übersicht: Kapitel 2

Die "Bestandsaufnahme" enthält grundlegende Unterscheidungen und Qualitätsmerkmale unter Berücksichtigung der deutschen Sprachtests für den Hochschulzugang (DSH und TestDaF). In Kapitel 2.1 werden testmethodische Klassifikationsmerkmale beschrieben und kritisch auf diese beiden Tests bezogen. In Kapitel 2.2 geht es um eine Analyse der Nützlichkeit beider Tests auf der Grundlage der Kriterien von Bachman und Palmer (1996).

Mit dem TestDaF hat die DSH eine attraktive Konkurrenz bekommen: Der TestDaF wurde zwischen 1999 und 2001 vom TestDaF-Institut entwickelt, er kann in Testzentren in der ganzen Welt an mehreren Terminen pro Jahr abgelegt werden, er ist testmethodisch auf der Höhe der Zeit. Die Zeugnisse werden von allen Hochschulen in Deutschland als Nachweis ausreichender Deutschkenntnisse anerkannt. Die kurze Geschichte des TestDaF ist eine Erfolgsgeschichte: Die kritischen Stimmen zum TestDaF resultierten in der Regel aus einer Abwehr von Änderungen und Ängsten vor Kompetenzverlust. Doch der Erfolg, messbar in stark steigenden Teilnehmerzahlen und einer innovativen Begleitforschung, gab dem TestDaF Recht und ließ Kritiker verstummen.

Die DSH ist aus der "Prüfung zum Nachweis deutscher Sprachkenntnisse" (PNdS) hervorgegangen. Sie kann an fast allen Hochschulen in Deutschland abgelegt werden. Sie wird von Sprachlehrern vor Ort erstellt und durchgeführt.

Es gibt mehr Trennendes als Gemeinsames zwischen den Prüfungen: Gemeinsam ist, dass beide Prüfungen als Nachweis der deutschen Sprachkenntnisse für ein Studium an den Hochschulen in Deutschland anerkannt werden. Gemeinsam ist ihnen auch, dass fast überall eine Gebühr für die Teilnahme verlangt wird. Für den TestDaF beträgt die Prüfungsgebühr um 100 Euro. Die Gebühren für die DSH variieren stark. Unterschiede bestehen nicht nur im unterschiedlichen Umgang mit der Grammatik und dem Fachbezug, den Schwerpunktthemen dieser Arbeit. Unterschiede bestehen auch im Grad der Standardisierung, in der Testfunktion, in der Bezugsgröße, in Merkmalen der Prüfungsteile, im Ergebnisausweis. Unterschiede zwischen beiden Tests werden anhand von testtheoretischen Merkmalen in Kapitel 2.1 herausgearbeitet. Unterschiede zwischen beiden Tests werden auch deutlich, wenn man beide Tests nach Kriterien zur Nützlichkeit von Sprachtests analysiert (Kapitel 2.2).

2.1. Klassifikation von Sprachtests für den Hochschulzugang

Übersicht: Kapitel 2.1

Thema dieses Kapitels sind Unterscheidungen, die für Sprachtests für den Hochschulzugang relevant sind. Unterschieden wird zwischen standardisierten und nicht-standardisierten Tests, es werden Sprachtests mit unterschiedlichen Funktionen vorgestellt und diskutiert (Lernfortschrittstests, Kursabschlusstests, Feststellungsprüfungen, Eignungstests, Einstufungstests, Zulassungsprüfungen und Diagnosetests), normorientierte und kriteriumsorientierte Sprachtests erörtert, Performanz- von Kompetenztests unterschieden sowie Modelle von Sprachkompetenz vorgestellt. Diese Klassifikationsmerkmale werden auf die DSH und den TestDaF bezogen. Neben der Darstellung und Anwendung der Klassifikationsmerkmale ist die kritische Würdigung der DSH und des TestDaF das Ziel des Kapitels.

Von deutschen Hochschulen werden zwei Deutschprüfungen als Nachweis für ausreichende Sprachkenntnisse für das Studium anerkannt, die "Deutsche Sprachprüfung für den Hochschulzugang" (DSH) und der "Test Deutsch als Fremdsprache für Studienbewerber" (TestDaF). Leisten sie das Gleiche? Es gibt eine Reihe von testmethodischen Unterschieden, wie aus der folgenden Beschreibung hervorgeht.

Standardisierte und nichtstandardisierte Tests

Man kann zwischen standardisierten und nichtstandardisierten Tests unterscheiden. Lienert und Raatz erläutern:

> *Standardisierte Tests* müssen wissenschaftlich entwickelt, hinsichtlich der wichtigsten Gütekriterien untersucht und unter Standardbedingungen durchführbar und normiert sein. Im anderen Falle handelt es sich um *nichtstandardisierte* oder *informelle* Tests, wie sie Psychologen und Lehrer gewissermaßen für den Hausgebrauch benützen und auswerten (1994: 14, Hervorhebungen im Original).

Folgt man der Definition von Lienert und Raatz, nach der es nur standardisierte oder nichtstandardisierte Tests gibt, so ist die DSH zu den nichtstandardisierten Tests zu zählen. Die Rahmenordnung für die DSH wurde von der Hochschulrektorenkonferenz 1995 in Kraft gesetzt und inzwischen von fast allen Hochschulen in Deutschland umgesetzt (HRK, 2000). Im Auftrag des "Fachverbandes Deutsch als Fremdsprache" wurde ein Handbuch für Prüferinnen und Prüfer herausgegeben, das die Rahmenordnung durch Anwendungsbeispiele und Hinweise zur Durchführung konkretisiert (FaDaF, 2001). Inzwischen liegt eine weitere Prüfungsordnung mit Aussagen zum TestDaF, zur DSH und zum Prüfungsteil Deutsch der Feststellungsprüfung vor (HRK/KMK, 2004).

Angesichts der fehlenden Standardisierung der DSH bezweifelt Alderson (2002b), dass die DSH überhaupt als Test anzusehen ist. Seiner Meinung nach ist es angemessener, von einem Prüfungsgerüst (*framework*) zu sprechen. Nach Lienert und Raatz eignet sich eine nichtstandardisierte Prüfung nur "für den Hausgebrauch" (s. o.). Das Todesurteil für die DSH? Perlmann-Balme hält eine geringe Standardisierung im Falle der DSH für unproblematisch:

> Die Frage von Abweichungen im Schwierigkeitsgrad bei abweichenden Prüfungsformaten von Universität zu Universität stellt sich insofern nicht als Problem, als jede Universität zugleich Prüfungsmacher und Endabnehmer der Zeugnisse, also zugleich prüfende und anerkennende Institution ist (Perlmann-Balme, 2001: 1000-1001).

Die Nähe der DSH zu den Kandidaten und den Anwendern ist sicherlich ein Vorteil der DSH. Die Schwierigkeiten, die sich aus der fehlenden Standardisierung ergeben, können jedoch nicht geleugnet werden: Erstens erkennen fast alle Hochschulen in Deutschland die DSH von anderen Institutionen als Nachweis der Deutschkenntnisse für ein Studium an, die prüfende Institution ist also mitnichten zwangsläufig auch die anerkennende Institution. Zweitens hat die Prüfung gewichtige Konsequenzen für die Kandidaten. Diese haben ein Recht auf eine Prüfung, die verlässliche Messergebnisse liefert. Auch innerhalb der jeweiligen Hochschule sollten die Institute, an denen die Prüfungskandidaten ein Studium aufnehmen, sich auf einen bestimmten Standard verlassen können. Die fehlende Standardisierung und die daraus resultierenden Probleme waren ein Anlass für die Entwicklung des TestDaF. Der Aspekt der Standardisierung stand beim TestDaF im Mittelpunkt der Testentwicklung (siehe "Reliabilität", Kapitel 2.2).

Funktionen von Sprachtests

Für Sprachtests gilt: Am Anfang steht die Funktion. Die Funktion bestimmt die Inhalte und auch die Fertigkeiten, die im Test geprüft werden. Millman und Greene raten Testentwicklern, sich unbedingte Klarheit über die Funktion zu verschaffen:

> The first and most important step in educational test development is to delineate the purpose of the test or the nature of the inferences intended from test scores. A clear statement of purpose provides the test developer with an overall framework for test specification and for item development, tryout, and review. A clear statement of test purpose also contributes significantly to appropriate test use in practical contexts (Millman/Greene, 1989: 335).

Doch die Klarheit existiert in der Praxis nur selten: Millman und Greene deuten in diesem Zitat bereits an, dass Tests nicht immer ihrer Funktion gemäß eingesetzt werden.

Das Thema Funktionen von Sprachtests verlangt zwei Blickrichtungen: Zunächst gilt es zu erläutern, unter welchen Voraussetzungen Tests für bestimmte Zwecke eingesetzt werden können und welche Auswirkungen bestimmte Verwendungszusammenhänge haben. Für Sprachtests für den Hochschulzugang ist zweitens zu klären, wie die Funktion, den Nachweis von Sprachkenntnissen für den Hochschulzugang, umgesetzt wird. Die Funktion von Sprachtests ist nach Chapelle und Read (1996; zit. n. Chapelle, 2001: 101-102) durch drei Faktoren zu bestimmen: die beabsichtigten Interpretationen der Ergebnisse, der beabsichtigte Umgang damit und die beabsichtigten Auswirkungen des Tests. Eine Typologie von Sprachtests nach ihrer Funktion existiert nicht, da in unterschiedlichen Einsatzbereichen eine Vielzahl von Funktionen vorkommt, welche sich nicht genau abgrenzen lassen.

Folgende Funktionen werden häufig unterschieden (Tabelle 2): Erstens Sprachtests, die einen Bezug zu einem Curriculum haben. Bei Lernfortschritts- bzw. Kursabschlusstests (*progress tests*, *exit tests* sowie allgemein *achievement tests*) werden die Ergebnisse der Kandidaten mit Blick auf einen bestimmten Lernstoff interpretiert. Zweitens Sprachtests, die sich auf ein sprachliches Konstrukt beziehen. Feststellungsprüfungen (*proficiency tests*) sollen Informationen über den Sprachstand der Kandidaten bereitstellen. Drittens Sprachtests, die sich auf die Zukunft beziehen. Eignungstests (*aptitude tests*) sollen beispielsweise die Erfolgsaussichten beim Sprachenlernen vorhersagen. Weitere Unterscheidungen nach der Funktion sind Diagnosetests (*diagnostic tests*), Einstufungstests (*placement tests*) sowie Zulassungsprüfungen (*selection tests* bzw. *gatekeeping tests*). Hier gibt es Überschneidungen und graduelle Unterschiede: Ein Ein-

stufungstest könnte beispielsweise auch eine Feststellungsprüfung sein (obwohl letztere in der Regel umfassender sind), ein Lernfortschrittstest könnte auch ein eingeschränkter Diagnosetest sein usw. (Brindley, 1990; Davies *et al*., 1999; Millman/Greene, 1989).

Tabelle 2: Sprachtests mit unterschiedlichen Funktionen

Sprachtests mit Bezug zu einem Curriculum:
Lernfortschrittstests (*progress tests, achievement tests*) Kursabschlusstests (*exit tests, achievement tests*)
Sprachtests mit Bezug zu einem sprachlichen Konstrukt:
Feststellungsprüfungen (*proficiency tests*)
Sprachtests mit Bezug zur Zukunft:
Eignungstests (*aptitude tests*) Einstufungstests (*placement tests*) Zulassungsprüfungen (*selection tests* bzw. *gatekeeping tests*)
Diagnosetests (*diagnostic tests*)

Die Validität eines Tests, welche nur mit Blick auf eine bestimmte Verwendung bestimmt werden kann, hängt eng mit der Funktion des Tests zusammen (Bachman, 1990; Davies, 1988; 1990; Messick, 1989; Shepard, 1993). Traditionell geht man davon aus, dass bei unterschiedlichen Testfunktionen jeweils andere Aspekte der Validität eine Bedeutung erlangen:

- bei Zulassungsprüfungen spielt die kriterienbezogene Validität die größte Rolle (d. h. wird das Testkriterium erfasst?),
- bei Kursabschlusstests die Inhaltsvalidität (d. h. wird der Kursinhalt – Sprache, Inhalt, Fertigkeiten – erfasst?),
- bei Feststellungsprüfungen die Übereinstimmungs- und prädiktive Validität (d. h. sind die Testergebnisse vergleichbar mit denen anderer Tests? Liefert der Test verlässliche Informationen über zukünftige Leistungen?),
- Augenscheinvalidität (halten Laien den Test für valide?) und Konstruktvalidität (hier verstanden als Bezug des Tests zu einem Modell der Sprachkompetenz) spielen bei jedem Test eine Rolle.

Diese Zuordnungen können dazu dienen, Eigenschaften von Tests mit einer bestimmten Funktion pointiert aufzuzeigen, auch wenn es in der Praxis sicherlich sinnvoller ist, möglichst viele Argumente für oder gegen die Validität eines Tests zu erfassen (Alderson, 2002a: 24-26).

Die Funktion eines Sprachtests lässt sich nicht immer eindeutig bestimmen. Die Unsicherheiten bei der Funktionsbestimmung liegen nicht nur an den Überschneidungen zwischen einzelnen Kategorien, sondern auch an unterschiedlichen Interessen der Testersteller, Testanwender und der Testteilnehmer. Die standardisierten Sprachtests für den Hochschulzugang wie TOEFL, IELTS sowie TestDaF sind, so werde ich argumentieren, Feststellungsprüfungen. Eine unanfechtbare Zuordnung ist dies nicht, andere Argumentationen sind denkbar. Bei der DSH sind andere Zuordnungen zutreffender: So scheint die Auswahl die vorrangige Funktion zu sein (Zulassungsprüfung). Mit der neuen Rahmenordnung wird aber auch bei der DSH der Charakter einer Feststellungsprüfung betont. Gründe für diese Einschätzungen und Besonderheiten der einzelnen Testfunktionen werden im Folgenden diskutiert.

Bei **Zulassungsprüfungen** geht es nicht nur um die Zulassung im engeren Sinne, also beispielsweise um die Zulassung zum Studium, sondern allgemein um die Auswahl von Teilnehmern aus einer größeren Gruppe (Chalhoub-Deville/Turner, 2000). Bei Zulassungsprüfungen kommt dem Schwellenwert oder dem kritischen Wert, welcher die Grenze zwischen Bestehen und Nicht-Bestehen darstellt, eine große Bedeutung zu. Wenn das Testergebnis weiter differenziert wird und diese Differenzierung nicht nur über Annahme oder Ablehnung entscheiden, sondern auch darüber, in welchen Kurs der Kandidat aufgenommen werden soll und welche Fertigkeiten besonders trainiert werden sollen, liegt eher ein Einstufungs- oder Diagnosetest vor (siehe unten). Das ist bei Sprachtests für den Hochschulzugang nicht der Fall.

Davies *et al*. (1999: 66) verweisen auf die besonderen Schwierigkeiten bei der Festlegung von Schwellenwerten: Normalerweise soll der Schwellenwert oberhalb der Ergebnisse von Testteilnehmern liegen, deren Leistungen nicht ausreichen. Was geschieht jedoch bei den Testteilnehmern, deren Testergebnis genau unter dem kritischen Wert liegt? Kann man wirklich so genau messen, dass sie auszuschließen sind oder sollte man aufgrund des Messfehlers eine gewisse Kulanz zeigen? Oder sollte man Testteilnehmer, deren Ergebnisse genau über dem kritischen Wert liegen, aufgrund von

Zweifeln an der Testgenauigkeit nicht mehr auswählen? In Zulassungsprüfungen kommt der kriterienbezogenen Validität daher eine große Bedeutung zu. Die kriterienbezogene Validität kann bestimmt werden über Korrelationen mit dem Kriterium oder mit einem Test, dessen kriterienbezogene Validität bekannt ist oder – und das dürfte bei der DSH der Fall sein – über die Meinung eines Experten (Davies *et al.*, 1999: 37-38; Grotjahn, 2000a: 313; Messick, 1989). Das Kriterium von Sprachtests ist eine bestimmte Qualität oder ein bestimmtes Niveau der Sprachkenntnisse. Im Fall von Sprachtests für den Hochschulzugang stellt die Prüfung der kriterienbezogenen Validität ein Problem dar, denn das Außenkriterium, an dem der Test gemessen werden soll, ist nicht genau zu bestimmen bzw. auch den Experten nicht genau bekannt.

Es bleibt abzuwarten, ob man die drei Ergebnisklassen der DSH als differenzierten Ausweis für ein bestimmtes Sprachniveau ansehen wird oder ob sie weiterhin allein unter dem Gesichtspunkt "bestanden – nicht bestanden" interpretiert werden. Da die inhaltliche Beschreibung der Stufen ungenau ist, wäre es verwunderlich, wenn sich die Ergebnisklassen als eine sinnvolle Beschreibung von Niveaustufen erweisen würden. Wahrscheinlicher ist, dass das Ergebnis einer DSH weiterhin allein unter dem Gesichtspunkt "bestanden – nicht bestanden" betrachtet wird. Für die Ergebnisklasse DSH-3 ist ohnehin keine Verwendung vorgesehen. Allein bei der DSH-1 könnte ein differenzierter Umgang erfolgen, da es als "beschränkt bestanden" gilt. Alles in allem erscheint es angebracht, die DSH weiterhin als Zulassungsprüfung anzusehen, bei der die kriterienbezogene Validität nicht geklärt ist (siehe folgender Abschnitt: "Normorientierte und kriteriumsorientierte Sprachtests").

In gewisser Hinsicht weist auch der TestDaF Eigenschaften einer Zulassungsprüfung auf. Zwar ist er aufgrund seines Bezugs auf ein theoretisches Konstrukt (Sprachkompetenz auf einem bestimmten Niveau) von der Konzeption her eine Feststellungsprüfung. Jedoch wird der TestDaF von Anwendern als Zulassungsprüfung genutzt. So können ausländische Studienbewerber die sprachlichen Voraussetzungen für ein Studium an der Fachhochschule Konstanz nachweisen, wenn sie im TestDaF im Durchschnitt die TDN 4 erzielen. Ähnlich verhält es sich beim *Test of English as a Foreign Language* (TOEFL): Die Hochschulen legen den Umgang mit den Ergebnissen fest. Für die Universitäten in Stanford und Yale gilt:

> Applicants taking the computer-based TOEFL are required to have a total score in the range of 230-300 to begin graduate study at Stanford. For doctoral programs in all fields and master's programs in

> Humanities, Social Sciences, and Education, a minimum total score in the range of 250-300 is required (Internetseite der *Stanford University*).
>
> ... A minimum score of 600 is required on the paper-based TOEFL; a minimum score of 250 is required on the computer-based TOEFL (Internetseite der *Yale University*).

Diese Aussagen werden von den Anwendern getroffen, nicht von den Testinstituten, welche die Aussage "bestanden – nicht bestanden" vermeiden, weil sie den Nutzen eines zentralen Sprachtests nicht allzu sehr einschränken möchten und weil sie sich nicht anmaßen, Zulassungsentscheidungen zu treffen, mit denen sie nicht direkt zu tun haben. Die Konstrukteure von standardisierten Sprachtests können und wollen den Abnehmern ihrer Prüfung nicht die Aufgabe abnehmen, das Niveau der Sprachkenntnisse festzulegen, denn die Kriterien, nach denen die Auswahl vorgenommen werden soll, unterscheiden sich.

Ich möchte als Zwischenergebnis festhalten: Die standardisierten Sprachprüfungen sind von den Testerstellern nicht als Zulassungsprüfung angelegt, die Abnehmer setzen sie jedoch als Zulassungsprüfung ein. Für die Testinstitute steht bei der Konzeption der Tests daher die Konstruktvalidität im Mittelpunkt, nicht die Ermittlung eines Schwellenwerts. Für die Anwender bekommt die Ermittlung des Schwellenwerts und damit die kriteriumsorientierte Validität eine größere Bedeutung. Die Anwender orientieren sich an den Ergebnisklassen der Prüfung, mit deren Hilfe sie aufgrund von Erfahrungen, Vereinbarungen oder Empfehlungen einen Schwellenwert festlegen. Bei der DSH verhält es sich bislang etwas anders. Bei ihr ist die Auswahl eine zentrale Funktion, eine Nähe zu Zulassungsprüfungen ist daher nicht von der Hand zu weisen. Ob sich dies mit dem differenzierten Ergebnisausweis in drei (bzw. vier) Ergebnisklassen ändert, bleibt abzuwarten.

Lernfortschritts- oder Kursabschlusstests beziehen sich auf einen Lernprozess. Vor allem bei der DSH stellt sich die Frage, ob sie wegen ihrer Nähe zur Prüfungsvorbereitung häufig die Funktion eines Kursabschlusstests annimmt. Auch bei den standardisierten Sprachtests für den Hochschulzugang ist es nicht ausgeschlossen, dass sie von den Abnehmern wie Kursabschlusstests behandelt werden, da den Prüfungen oft Vorbereitungskurse vorausgehen. Daher möchte ich auf die Besonderheiten von Lernfortschritts- und Kursabschlusstests kurz eingehen.

Für Kursabschlusstests und Lernfortschrittstests gilt: Geprüft wird, was unterrichtet wurde. Kursabschlusstests schauen ähnlich wie Lernfortschrittstests zurück auf einen

Lernweg, sie beziehen sich jedoch nicht auf ein bestimmtes Kurssegment, sondern auf den gesamten Stoff des Kurses (Alderson, 2002a; Davies, 1990; Davies *et al*., 1999: 2; Hughes, 1988: 37; Perlmann-Balme, 2001: 1003). Sie werden in der Regel dezentral von Prüferinnen und Prüfern erstellt, welche die Testteilnehmer selbst unterrichtet haben. Normalerweise sind die Konsequenzen einzelner Lernfortschrittstests nicht sehr weit reichend, in einigen Situationen können Kursabschlusstests dennoch schwer wiegende Konsequenzen haben: Sie können z. B. darüber entscheiden, ob ein Kurs wiederholt werden muss. Wenn zu Beginn der nächsten Stufe keine Zulassungsprüfung steht, werden die Ergebnisse aus Kursabschlusstests häufig auch als Selektionsinstrument genutzt. Das ist nur legitim, wenn sich Inhalte und Anforderungsniveau der Kurse, an denen die Bewerber teilgenommen haben, nicht unterscheiden – was allerdings in der Regel nicht zutrifft. Vom Standpunkt der Testmethodik sind daher Feststellungsprüfungen, Zulassungsprüfungen oder mindestens zentrale Kursabschlusstests dezentralen Kursabschlusstests vorzuziehen.

Gibt es einen Platz für eine Prüfung, welche ausschließlich Auskunft über die Leistung in einem Kurs gibt? Der Nutzen einer solchen Prüfung ist beschränkt: Sie ist für Kursteilnehmer, für Lehrkräfte, möglicherweise für die Institutsleitung oder Schulaufsicht als Rückmeldung von Interesse. Für die Verwendung in anderen Zusammenhängen ist eine Kursabschlussprüfung nur sinnvoll, wenn ein klar definiertes Kursziel verfolgt wurde (Hughes, 1988).

Anders als bei Feststellungsprüfungen stellt die Bestimmung der Validität von Kursabschlusstests kein allzu großes Problem dar. Zum Tragen kommt bei Kursabschlusstests vor allem die Inhaltsvalidität, die das Ausmaß angibt, in dem der Test den Lernstoff (Inhalte, Fertigkeiten, Sprache) erfasst (Grotjahn, 2000a: 312). Die Validität eines Kursabschlusstests hängt von der Auswahl der Themen und der Sprache ab. Wenn diese Auswahl repräsentativ für den Kurs ist, ist der Test valide. Schwierigkeiten mit der Validität von Kursabschlusstests ergeben sich, wenn die Ergebnisse für andere Zwecke verwendet werden. Bei der Bestimmung der Inhaltsvalidität ist man auf das Urteil von Experten angewiesen, unter Umständen sind Korrelationen mit Lernfortschrittstests möglich. Das Testergebnis eines Kursabschlusstests sagt nichts darüber aus, ob das Niveau und die Qualität des Kurses angemessen waren oder ob die Themen und die Sprache in irgendeiner Weise relevant waren. Es ist denkbar, dass der Kurs sein Ziel, bei-

spielsweise die sprachliche Vorbereitung auf ein Studium, durch unangemessene Inhalte nicht erreicht hat. Die Kursabschlussprüfung könnte unabhängig von der Eignung der Teilnehmer zum Studium valide sein. Kursabschlusstests mit schwer wiegenden Konsequenzen haben wohl am ehesten einen Einfluss auf die Lehr- und Lernprozesse. Es besteht die Tendenz, dass aus einem Sprachkurs ein Kurs zur Prüfungsvorbereitung wird.

Standardisierte Sprachtests für den Hochschulzugang wie der TestDaF sind keine Kursabschlusstests. Das ist sachgemäß, denn Hochschulen, an denen die ausländischen Studienbewerber studieren werden, interessieren sich nicht dafür, ob irgendein Sprachkurs mit Erfolg bewältigt wurde. Von Interesse ist allein, ob ausreichende Sprachkenntnisse für ein Studium vorhanden sind. Wie diese erworben wurden, ist unerheblich. Die Vorbereitung ausländischer Studierender ist sehr uneinheitlich. Es gibt kein einheitliches Kurssystem, mit dem auf ein Studium in Deutschland vorbereitet wird. Wichtig ist vielmehr die in die Zukunft gerichtete Information, ob ausreichende Sprachkenntnisse für ein Studium vorliegen.

Auch bei der DSH sollte es sich eigentlich nicht um einen Kursabschlusstest handeln. Durch die Nähe zu den prüfenden Instituten ist jedoch zu erwarten, dass die DSH häufig den Charakter eines Kursabschlusstests annimmt. Dies wird besonders am DSH-Grammatiktest deutlich, der Thema der Kapitel 3 und 4 ist. Auch am Umgang mit dem Fachbezug – dem Thema der Kapitel 5 und 6 –, wird die Nähe zu einem bestimmten Lernplan deutlich. In der Rahmenordnung für die DSH wird zur Textauswahl für den Prüfungsteil Leseverstehen ausgeführt:

> Es soll ein weitgehend authentischer, studienbezogener und wissenschaftsorientierter Text vorgelegt werden, der keine Fachkenntnisse voraussetzt, ggf. nur solche, die Gegenstand eines vorangegangenen fachspezifisch orientierten Unterrichts waren (HRK, 2000; in der überarbeiteten Rahmenordnung unverändert, vgl. HRK/KMK, 2004).

Die Passage verdeutlicht, dass ein enger Zusammenhang zwischen der Prüfungsvorbereitung und der DSH als selbstverständlich angesehen wird. Ich halte es aus den oben erwähnten Gründen nicht für sinnvoll, wenn ein Kursabschlusstest als Sprachtests für den Hochschulzugang verwendet wird.

Ähnlich verhält es sich übrigens mit Prüfungen des Goethe-Instituts. Das "Kleine Deutsche Sprachdiplom" (KDS) enthält beispielsweise einen Prüfungsteil "Fragen zur Lektüre". Diesen Prüfungsteil kann nur beantworten, wer das Buch auch gelesen hat. Diese Prüfung legt eine bestimmte Vorbereitung also nahe, bestimmte Inhalte müssen

vorher erarbeitet werden. Für die Testteilnehmer ist es dann angezeigt, an Kursen teilzunehmen, welche von erfahrenen Lehrkräften geleitet werden, die den Umgang mit der Lektüre in der Prüfung gut einschätzen können.

Feststellungsprüfungen sind in der Regel standardisierte Sprachtests, die sich nicht an dem Lernstoff eines bestimmten Kurses orientieren. Ihre Aufgabe ist eine Beschreibung der gegenwärtigen sprachlichen Fähigkeiten unabhängig vom Lernweg (Alderson, 1988; Davies, 1990; Perlmann-Balme, 2001). Da Feststellungsprüfungen ein mehr oder weniger umfassendes Bild der Sprachkenntnisse bieten sollen, sind sie recht umfangreich und zeitaufwändig. Häufig werden die Ergebnisse von Feststellungsprüfungen mit dem Ziel interpretiert, Aussagen über die Angemessenheit der Sprachkenntnisse für eine bestimmte Aufgabe zu gewinnen: den Nachweis von ausreichenden Sprachkenntnissen, um einen Kurs in einem bestimmten Fach zu belegen, um als Fremdenführer Stadtführungen durchzuführen oder eben um ein Hochschulstudium ohne größere Sprachschwierigkeiten absolvieren zu können. Die Aussage von Feststellungsprüfungen richtet sich dann in die Zukunft. Die Ergebnisse von Feststellungsprüfungen sind wegen des Bekanntheitsgrads dieser Prüfungen und ihrer Unabhängigkeit von einem Curriculum auch für Außenstehende interpretierbar. Das ist bei Kursabschlusstests nicht immer der Fall.

Bei Feststellungsprüfungen kommen ihrer Funktion gemäß zwei Aspekte der Validität zum Tragen: die prädiktive Validität und die Übereinstimmungsvalidität. Beide werden über Korrelationen mit einem Außenkriterium ermittelt. Die prädiktive Validität bezieht sich auf die Übereinstimmung der Testleistung mit einem Kriterium, das sich auf eine Sprachverwendungssituation in der Zukunft bezieht. Die Übereinstimmungsvalidität bezieht sich auf die Übereinstimmung mit Leistungen in einem anderen Test (Davies, 1988; Messick, 1988). Darüber hinaus spielt das übergreifende Konzept der Konstruktvalidität bei Feststellungsprüfungen eine besondere Rolle. Die Aussage von (sprachlichen) Feststellungsprüfungen bezieht sich auf ein theoretisches Konstrukt, häufig Sprachkompetenz auf einem bestimmten Niveau. Argumente für eine hohe Konstruktvalidität lassen sich nur finden, wenn der Test methodisch abgesichert ist und auf einem soliden theoretischen Fundament steht. Der Test soll möglichst alle Aspekte aufweisen, welche für das Konstrukt repräsentativ sind ("*construct underrepresentation*"), und möglichst wenig Aspekte aufweisen, welche von dem Konstrukt ablenken oder welche

für das Konstrukt nicht relevant sind ("*construct irrelevance*"). Dies ist für Feststellungsprüfungen von Bedeutung, weil diese nicht auf einen Lehrplan zurückgreifen können und weil sie sich nicht auf eine bestimmte Auswahlentscheidung oder Kurszuordnung beziehen (Alderson, 2002a: 26-27).

Aus der Ungenauigkeit von Konstrukten wie "allgemeine Sprachkompetenz" ergeben sich Schwierigkeiten für Feststellungsprüfungen, nicht nur mit Blick auf die Konstruktvalidität, sondern beispielsweise auch für die Auswahl der Inhalte. Bei der Auswahl der Inhalte für eine Feststellungsprüfung kann man sich nicht wie bei Kursabschlussprüfungen an einem bestimmten Lehrplan oder an einem konkreten Unterrichtsgeschehen orientieren. Es wurde bereits darauf hingewiesen, dass die Vorbereitungen auf Feststellungsprüfungen sehr unterschiedlich sind, Anhaltspunkte für die Testerstellung lassen sich daraus nicht gewinnen; es wäre auch nicht konform mit dem Gedanken einer Feststellungsprüfung. Testersteller von Feststellungsprüfungen müssen die Inhalte ausgehend von einer Sprachbedarfsanalyse des Testkonstrukts selbst entwickeln. In der Praxis kommt es ungewollt zur Entstehung von speziellen "Prüfungscurricula" (Alderson, 1988). Dies wird an einer Zusammenstellung von Prüfungsinhalten deutlich.

Die Entstehung eines Prüfungscurriculums kann in der Praxis zu einem veränderten Umgang mit Feststellungsprüfungen führen, wie Alderson (2000b: 23-24) am Beispiel der Prüfung "*First Certificate in English*" (FCE) erläutert. Von dem Testinstitut, UCLES, ist diese Prüfung als Feststellungsprüfung intendiert. Was das FCE feststellt, mit anderen Worten, welchen Konstrukts sie sich bedient, ist jedoch unklar. Wahrscheinlich liegt ihre Leistung eben darin, Sprachkenntnisse auf dem Niveau des FCE-Tests ausweisen zu können! Intern wird die Validität möglicherweise durch einen Vergleich mit leichteren und anspruchsvolleren Tests aus dem eigenen Haus oder durch eine Korrelation der Ergebnisse mit Tests, welche Sprachkenntnisse auf gleichem Niveau ausweisen sollen, erzielt. Da die Prüfung anerkannt und verbreitet ist, gibt es eine Vielzahl von Sprachkursen, welche auf den Test vorbereiten. Es werden Unterrichtsmaterialien eingesetzt, welche mit Blick auf die Spezifikationen des FCE verfasst wurden. Möglicherweise erzielt das FCE eine hohe Inhaltsvalidität, so mutmaßt Alderson nur halb im Spaß, insbesondere durch einen Abgleich des Tests mit den Materialien zur Testvorbereitung. Diese werden in großer Zahl auch von UCLES publiziert (Alderson, 2000b). Wenn das gewissenhafte Durcharbeiten von Lehrbüchern zum Be-

stehen eines Sprachtests führt und wenn sich der Sprachtest an diesem Lehrbuch orientiert, dann handelt es sich weniger um eine Feststellungsprüfung als vielmehr um eine Kursabschlussprüfung.

Die Möglichkeit der Validierung über die Inhalte der Vorbereitung, wie sie im vorangegangenen Absatz etwas überzeichnet dargestellt wurde, bietet sich den Erstellern zentraler Sprachtests für den Hochschulzugang nur theoretisch. In der Realität ist von einer derartigen Praxis keineswegs auszugehen. Testteilnehmer, Autoren von prüfungsvorbereitendem Material und Lehrkräfte in prüfungsvorbereitenden Kursen haben jedoch ein starkes Interesse daran, einen Prüfungslehrplan zu konkretisieren und abzuarbeiten. Aus den veröffentlichten Prüfungen lässt sich ein Kanon aus Themen, Lexik, Strukturen usw. gewinnen.

Schließlich sollte zum Thema Einsatz von Feststellungsprüfung als Kursabschlusstest noch die Perspektive der Testteilnehmer in Erinnerung gerufen werden. Sie möchten sich vorbereiten auf eine Prüfung, erhalten von einer Feststellungsprüfung wegen der inhaltlichen Offenheit aber wenig konkrete Hinweise auf eine Vorbereitung. Kursabschlussprüfungen bieten Prüfungskandidaten mehr Sicherheit und Orientierung.

Standardisierte Sprachtests für den Hochschulzugang werden von den Testinstituten als Feststellungsprüfung konzipiert. Das gilt auch für den TestDaF. Das bedeutet jedoch nicht zwangsläufig, dass sie in der Praxis ebenfalls wie eine Feststellungsprüfung behandelt werden, vielmehr kann die Prüfung über das Abarbeiten eines Prüfungscurriculums wie eine Kursabschlussprüfung behandelt werden. Im Falle des TestDaF ist außerdem zu beobachten, dass die Abnehmer den Umgang mit den differenzierten Ergebnisklassen, die auf dem Zeugnis ausgewiesen werden, erst lernen müssen. Der nach sprachlichen Fertigkeiten differenzierte Ergebnisausweis bedient die von den Hochschulen erwartete Information, ob die Voraussetzungen für die Zulassung vorliegen, nicht direkt. Die Ergebnisse müssen interpretiert werden. Möglicherweise ist man beim TestDaF in der Differenzierung auch im oberen Leistungsspektrum über das Ziel hinausgeschossen.

Auch die DSH soll eine Feststellungsprüfung sein. Dies ist jedenfalls die Absicht der Hochschulrektorenkonferenz und des Fachverbandes Deutsch als Fremdsprache. Dies wurde an den überarbeiteten Richtlinien deutlich: Auch bei der DSH sollen die Ergeb-

nisse in den einzelnen Prüfungsteilen auf dem Zeugnis separat ausgewiesen werden. Auf diese Weise soll ein Bild der Sprachkenntnisse entstehen. Dies ist zu begrüßen, die Schwierigkeit liegen wiederum in der fehlenden Standardisierung.

Ich möchte das Thema Funktionen von Sprachtests nicht abschließen, ohne auf drei weitere Unterscheidungen hinzuweisen: Diagnosetests, Einstufungstests und Eignungstests.

Diagnosetests sollen Fremdsprachenlernern Hinweise über ihren Lernstand geben, über ihre Fähigkeiten und ihre Defizite (Hughes, 2003: 11-26). Mit diesen Informationen können Rückschlüsse über den Lernprozess gezogen sowie Empfehlungen für eine Organisation des weiteren Lernprozesses ausgesprochen werden. Ein C-Test (Erläuterung siehe Kapitel 1, Seite 8) wäre kein geeignetes Testinstrument für einen Diagnosetest, denn das Ergebnis des C-Tests ist ein bestimmter Wert, aus dem nicht hervorgeht, wie er zustande gekommen ist. Ein Test eignet sich für Aussagen über Stärken und Schwächen des Sprachstands, wenn das Ergebnis mittels qualitativer Deskriptoren ausgewiesen werden kann oder wenn Test-Rohwerte eine inhaltliche Interpretation zulassen.

Eine Diagnose der sprachlichen Stärken und Schwächen ist nicht das Hauptanliegen von Sprachtests für den Hochschulzugang. Nur in Ausnahmefällen wenden sich Testteilnehmer, welche beispielsweise die DSH nicht bestanden haben, an die Institute mit der Bitte um Informationen, welche sprachlichen Schwächen zum Nichtbestehen geführt haben könnten. Das Ergebnis des TestDaF enthält durch den Ausweis von Niveaustufen in den vier Prüfungsteilen einige Informationen, welche Rückschlüsse auf Stärken und Schwächen in bestimmten Fertigkeiten zulassen. Der TestDaF ist jedoch nicht als Diagnosetest angelegt und wäre als Diagnosetest nicht differenziert genug. Das gilt auch für die DSH.

Einstufungstests haben eine Ähnlichkeit zu Diagnosetests, jedoch verfolgen sie das konkrete Ziel der Kurszuweisung (Wall/Clapham/Alderson, 1994; Perlmann-Balme, 2001). Mit Hilfe von Einstufungstests werden Zuordnungen zu bestimmten Lerngruppen vorgenommen. Der Inhalt orientiert sich an dem Inhalt der jeweiligen Kurse, meistens in Form von Stichprobenkontrollen.

Schließlich sollen **Eignungstests** erwähnt werden. Im Zusammenhang mit Sprachtests prüfen sie die "Fähigkeit zum Erlernen einer Sprache" (Davies *et al*., 1999: 10-11). Sie beruhen auf einer Theorie von Sprache, daher spielt die Konstruktvalidität eine bedeutende Rolle. Ein Beispiel für einen Eignungstest ist der "*Modern Language Aptitude Test*" (MLAT), der 1959 in den USA von Carroll für den *Foreign Service* entwickelt wurde (Spolsky, 1995: 117-133). Die Themen des Tests waren Zahlengedächtnis, visuelles Gedächtnis, Zuordnung von Graphemen und Phonemen, Arbeit mit Synonymen und grammatisches Verständnis. Beim Einsatz des Tests wurden durchschnittliche Korrelationen von 0,5 zwischen dem Testergebnis vor Kursbeginn und einem Testergebnis nach Kursende beobachtet (Erläuterung von Korrelationen: Kapitel 4.2.1, Seite 125). Es ist anzunehmen, dass Erfolg in einem Sprachkurs nicht allein auf die Fähigkeit zum Spracherwerb zurückzuführen ist. Andere Faktoren, wie Motivation und Qualität des Unterrichts oder der Betreuung spielen sicherlich auch eine Rolle für den Lernerfolg. Möglicherweise besteht der Nutzen derartiger Tests vor allem darin, Prüfungskandidaten abzulehnen, die im Eignungstest ein niedriges Ergebnis haben. Dann hätten Eignungstests eine ähnliche Funktion wie Zulassungsprüfungen.

Zu den **Funktionen von Sprachtests für den Hochschulzugang** ist festzuhalten, dass der TestDaF als Feststellungsprüfung angelegt ist. Ziel ist die Beschreibung des Sprachstands unabhängig von einem Curriculum. Ob sich in der Prüfungsvorbereitung durch die Erwartungen der Lehrenden und der Testteilnehmer ein informeller Lehrplan bildet, bleibt abzuwarten. Die Testanwender werden den TestDaF auch wie eine Zulassungsprüfung verwenden, enthalten jedoch mehr Informationen als bei der DSH. Die DSH ist eher eine Zulassungsprüfung. Da die DSH nun ebenfalls die Ergebnisse in einzelnen Prüfungsteilen ausweisen soll, wird der Charakter einer Feststellungsprüfung stärker betont. In der Praxis wird die DSH häufig wie eine Kursabschlussprüfung eingesetzt, d. h. es dürfte ein konkreter Zusammenhang zwischen der Vorbereitung und dem Test bestehen. Prüfungskandidaten können sich dann gezielter vorbereiten, die Aussagekraft für Testanwender ist jedoch geringer.

Normorientierte und kriteriumsorientierte Tests

Sprachtests lassen sich nach Art der Bezugsgröße unterscheiden. Werden die Rohdaten in Beziehung gesetzt zu denen anderer Testteilnehmer oder mit einem sprachlichen Kriterium? Im ersten Fall ist der Test normorientiert, im zweiten kriteriumsorientiert. Diese Unterscheidung hat Auswirkungen auf die Testkonstruktion und -auswertung. Beide Vorgehensweisen kommen bei Sprachtests für den Hochschulzugang vor, Ziel sind jedoch Aussagen zu einem Kriterium.

Bei normorientierten Sprachtests werden die Ergebnisrohwerte in Beziehung gesetzt zu den Ergebnissen anderer Testteilnehmer. Da das Ziel die Differenzierung zwischen Testteilnehmern ist, wird nicht der Rohwert ausgewiesen, sondern beispielsweise ein Prozentrang mit Bezug auf die Gesamtgruppe oder mit Bezug auf eine größere Gesamtheit, eine Vergleichsgruppe (Davies, 1988). Brown und Hudson definieren normorientierte Tests wie folgt:

> Any test that is primarily designed to disperse the performances of students in a normal distribution based on their general abilities, or proficiencies, for purposes of categorizing the students into levels or comparing students' performances to the performances of the others who formed the normative group (Brown/Hudson, 2002: 2).

Normorientierte Tests werden eingesetzt, um zwischen Testteilnehmern zu differenzieren und nicht um bestimmte Fähigkeiten zu beschreiben. Normorientierte Tests benötigen Items, welche stark differenzieren (Brown, 1996; Brown/Hudson, 2002: 6-14; Gronlund, 1988: 14).

TOEFL ist ein normorientierter Sprachtest. Das Testergebnis, welches den Testteilnehmern mitgeteilt wird, ergibt sich nicht aus den addierten Rohwerten, sondern aus den gewichteten Werten. Dieses Verfahren führt wegen der großen Anzahl der Testteilnehmer dazu, dass gleiche Ergebniswerte auch einer vergleichbaren Sprachkompetenz entsprechen, unabhängig von dem Schwierigkeitsgrad der jeweiligen Testversion und damit auch unabhängig von den jeweiligen Rohwerten, welche vom spezifischen Schwierigkeitsgrad abhängen. Prüfungskandidaten und Testanwender erhalten verlässliche Differenzierungen zwischen Kandidaten mit Blick auf die gewichteten Sprachkenntnisse (Brown, 1996; Brown/Hudson, 2002). Eine inhaltliche Beschreibung des Leistungsstands ermöglicht der TOEFL nicht. Das Testinstitut veröffentlicht in regelmäßigen Abständen Hinweise zum Umgang mit den Ergebnissen (ETS, 2000a; 2003). Dort findet man Informationen über das Abschneiden von Testteilnehmern nach be-

stimmten Kriterien (Herkunftssprache, Herkunftsland, Ausbildung, Beruf usw.) sowie eine Liste mit Hochschulen und Informationen über den Umgang mit den Ergebnissen. Es gibt eine Reihe von Untersuchungen zur Interpretation der Ergebnisse, die sich jeweils nicht auf ein bestimmtes Niveau der Sprachkompetenz beziehen, sondern auf einen Wert, ein bestimmtes Testergebnis. Messner und Liu fassen die Ergebnisse einer Longitudinaluntersuchung zur Studienleistungen ausländischer Studierender wie folgt zusammen: "International students with TOEFL scores of above 550 have a high likelihood of achieving academic success"[1] (Messner/Liu, 1995: 39). Durch den hohen Bekanntheitsgrad und die hohe Reliabilität des TOEFL werden die Ergebnisse in der Praxis häufig wie eine Beschreibung individueller sprachlicher Fertigkeiten behandelt. Die Ergebnisse werden von den Abnehmern beispielsweise mit bestimmten Fertigkeitsstufen in Verbindung gesetzt, oder Testersteller richten ihre Tests gar am TOEFL aus, erheben den Wert damit zu einem Kriterium (Brown/Hudson: 2002: 31).

Bei kriteriumsorientierten Tests geht es nicht um einen Vergleich der Ergebnisse mit einer Referenzgruppe, sondern um die absoluten Ergebnisse mit Blick auf ein bestimmtes Kriterium (Davies, 1988). Brown und Hudsons Definition kriteriumsorientierter Tests lautet:

> A criterion-referenced test [...] is primarily designed to describe the performances of examinees in terms of the amount that they know of a specific domain of knowledge or set of objectives (Brown/Hudson, 2002: 5).

Man kann sprachlich noch genauer unterscheiden, beispielsweise findet man in der Literatur lehrziel- bzw. lernzielorientierte Tests als Untergruppe kriteriumsorientierter Tests beschrieben (Brown, 1996; Brown/Hudson, 2002: 3-5; Grotjahn, 2000a: 330). Die Beschreibung der individuellen Fertigkeiten ist typischerweise ein Anliegen kriteriumsorientierter Tests. Anders als bei normorientierten Tests, bei denen es um die Differenzierung der Testteilnehmer geht, ist es bei kriteriumsorientierten Tests möglich, dass alle Prüfungskandidaten das gleiche Ergebnis erzielen. So könnten alle Kandidaten 100 Prozent erzielen! In umfangreichen Prüfungen wird dies selten der Fall sein, aber bei einzelnen Items ist es schon möglich, dass alle Testteilnehmer die volle Punktzahl erzielen. Für einen normorientierten Test wäre dieses Item sinnlos, weil es nicht zwischen den Leistungen der Kandidaten differenziert, als Indiz für das Kriterium ist es in einem kriteriumsorientierten Test möglich.

[1] Mit der Einführung des computerbasierten Tests wurde die TOEFL-Skala verändert, so dass diese Aussage nicht mehr zutrifft (ETS, 2000b).

Sowohl Feststellungsprüfungen als auch Kursabschlussprüfungen können kriteriumsorientiert vorgehen. Da kriteriumsorientierte Tests Bezug nehmen auf bestimmte Fertigkeiten, haben sie häufig auch einen engen Bezug zu einem bestimmten Curriculum bzw. zum Unterricht. Die deutschen Sprachtests für den Hochschulzugang sind kriteriumsorientiert: Sie sollen ermitteln, ob die Kandidaten "ausreichende Sprachkenntnisse für das Studium" erreicht haben. Dies gilt insbesondere für den TestDaF, der sich an sprachlichen Niveaustufen orientiert (Arras/Grotjahn, 2002).

Die DSH ist ebenfalls kriteriumsorientiert, allerdings ist die kriterienbezogene Validität nicht gesichert:

- Erstens sind bei der DSH Änderungen im Sinne einer Normorientierung unter dem Eindruck von unerwarteten oder unerwünschten Ergebnissen denkbar: Würden die Ausrichter einer DSH alle Testteilnehmer durchfallen lassen, wenn keiner der Teilnehmer über ausreichende Deutschkenntnisse für die Aufnahme eines Studiums verfügt? Möglicherweise würden Zweifel aufkommen, möglicherweise würde man den besten der schlechten Teilnehmer doch bestehen lassen, möglicherweise würde man unter diesem Eindruck die nächste Prüfung auf einem niedrigeren Niveau ansetzen.

- Zweitens wird das Testkriterium nicht inhaltlich, sondern nur über willkürlich festgelegte Schwellenwerte definiert. Testmethodisch mehr als fragwürdig sind die Ausführungen in der DSH-Rahmenordnung zur Bestimmung des Schwellenwerts bzw. der Ergebnisklassen. In der "DSH-Musterprüfungsordnung" heißt es:

  > Das Gesamtergebnis der Prüfung […] wird festgestellt:
  > - als DSH-1, wenn sowohl in der schriftlichen als auch der mündlichen Prüfung mindestens 57 % der Anforderungen erfüllt wurden;
  > - als DSH-2, wenn sowohl in der schriftlichen als auch der mündlichen Prüfung mindestens 67 % der Anforderungen erfüllt wurden;
  > - als DSH-3, wenn sowohl in der schriftlichen als auch der mündlichen Prüfung mindestens 82 % der Anforderungen erfüllt wurden (HRK/KMK, 2004: DSH-Musterprüfungsordnung §5(7).

  Die Leistungsstufen werden inhaltlich auf dem Zeugnis beschrieben:

  > DSH-3: Besonders hohe Fähigkeit, …
  > DSH-2: Differenzierte Fähigkeit, …
  > DSH-1: Grundlegende Fähigkeit, …
  > wissenschaftssprachliche Strukturen:
  > typische wissenschaftssprachliche Formen zu verstehen und selbst anzuwenden: Satzbau, wissenschaftliche Terminologie und Wortbildung, Wortschatz und Ausdrucksformen in unterschiedlichen Anwendungsbereichen, wie referierende Darstellung, argumentative Darlegung, … (HRK/KMK, 2004: DSH-Musterzeugnis).

Eine Anlehnung an bestimmte Sprachniveaustufen fehlt. Möglich wäre z. B. die Verankerung mit den Niveaustufen des Europarats (Europarat/Rat für kulturelle Zusammenarbeit, 2001). Der inhaltliche Bezug erfolgt in einem Zirkelschluss auf die (willkürlichen) Schwellenwerte: "Das Prüfungszeugnis dokumentiert die mit einzelnen Ergebnissen nachgewiesenen sprachlichen Fähigkeiten" (HRK/KMK, 2004: § 3). Umso erstaunlicher ist die genaue Zahl, welche ein fein abgestuftes Messniveau suggeriert. Diese Anweisung berücksichtigt außerdem nicht, dass die Schwierigkeitsgrade einzelner Prüfungen stark voneinander abweichen können. Sie ist wohl eher dahingehend zu verstehen, dass der Test nicht zu leicht sein soll. Wenn beispielsweise eine Normalverteilung auf der Prozentskala vorliegt (mit einer Standardabweichung von 17 Prozent), würden nur etwa 15 Prozent aller Kandidaten DSH-2 bzw. DSH-3 erzielen (und damit die uneingeschränkte sprachliche Studierfähigkeit nachweisen). In der Praxis kommen die Testausrichter eigentlich nicht umhin, jeweils neu festzulegen, bei welcher Punktzahl "ausreichende Sprachkenntnisse für das Studium" vorliegen. Ich vermute, dass die Testausrichter einen Erfahrungswert verwenden, um dieses Problem zu lösen.

Dass auch über einen normorientierten Test hinreichende Informationen für den Hochschulzugang ermittelt werden können, ist mit dem TOEFL hinreichend belegt. Von den Nutzern können die Ergebnisse des TOEFL mit Blick auf das Kriterium durchaus als verlässlich interpretiert werden, damit ist er in dieser Hinsicht ein nützlicher Test. Freilich ist der TOEFL durch die hohen Teilnehmerzahlen testmethodisch abgesichert.

Performanz- und Kompetenztests, direkte und indirekte Tests

Die Unterscheidung der Sprachtester zwischen Performanz- und Kompetenztests geht zurück auf die Unterscheidung des Linguisten Chomsky zwischen Sprachgebrauch (*performance*) und Sprachkompetenz (*competence*):

> We thus make a fundamental distinction between *competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of language in concrete situations), (Chomsky, 1965: 4; Hervorhebungen im Original).

"Performanz" in Chomskys Sinne ähnelt dem Begriff *parole*, unter dem de Saussure die konkrete Realisierung von Sprache in Gebrauch versteht. Die Aufgabe für die Linguistik besteht nach Chomsky darin, aus der Sprachverwendung auf die Sprachkompetenz zu schließen.

> The problem for the linguist, as well as for the child learning the language, is to determine from the data of performance the underlying system or rules that has been mastered by the speaker-hearer and that he puts to use in actual performance. Hence, in the technical sense, linguistic theory is mentalistic, since it is concerned with discovering a mental reality underlying actual behaviour (Chomsky, 1965: 4).

In dieser Kernfrage der Linguistik liegt auch das Kernproblem beim Umgang mit Sprachtests. Skehan (1984) formuliert mit den folgenden drei Fragen die Problematik aus Sicht der Sprachtester: Wie kann erstens von einer beobachteten sprachlichen Leistung auf sprachliche Fertigkeiten geschlossen werden? Zweitens: Wie kann von einer beobachteten sprachlichen Fertigkeit auf eine sprachliche Leistung geschlossen werden? Und drittens: Wie kann von einem konkreten inhaltlichen Zusammenhang auf einen anderen geschlossen werden? Bei Sprachtests sind zwei Umgangsweisen mit der Problematik der Interpretation von Testergebnissen anzutreffen:

Der erste Umgang schafft sich ein Modell, welches einen Zusammenhang zwischen sprachlicher Leistung und den zugrunde liegenden Fähigkeiten herstellt. Dieser Ansatz setzt eine genaue Untersuchung sprachlicher Leistungen und eine begründete Ableitung von bestimmten Fähigkeiten voraus, damit eine zutreffende und nachvollziehbare Beziehung zwischen beiden beschrieben werden kann. Das Modell sollte außerdem auf alle Sprecher in mehr oder weniger ähnlicher Art und Weise zutreffen. Bei Sprachtests, welche auf einem derartigen Modell beruhen, handelt es sich um Kompetenztests.

Der zweite Umgang bedient sich einer Abkürzung: Indem der Sprachtest die reale Sprachverwendungssituation, über die mit dem Test Aussagen gewonnen werden sollen,

möglichst realitätsnah abbildet und die Testteilnehmer mit dieser Situation konfrontiert, kann mehr oder weniger direkt auf Leistungen in eben dieser Situation geschlossen werden. Auf die "Aufdeckung der mentalen Realität, die dem aktuellen Verhalten zugrunde liegt," wird verzichtet. Dies ist die Vorgehensweise direkter Performanztests. Nicht gelöst wird jedoch das dritte der von Skehan skizzierten Probleme: das der Übertragbarkeit. Aussagen über mögliche Leistungen in anderen Sprachverwendungssituationen lässt ein derartiger Test in der Regel nicht zu.

In einem sprachlichen **Performanztest** (*performance test*) wird die Fähigkeit gemessen, bestimmte Sprachverwendungssituationen zu meistern. Das Kriterium eines (direkten) Performanztests entspricht weitgehend der tatsächlichen Sprachverwendungssituation. In Performanztests geht es um Sprache in Aktion; im Zentrum steht, was mit Sprache in einer bestimmten Verwendungssituation gemacht wird (Davies *et al.*, 1999; Grotjahn, 2000a; McNamara, 1996). McNamara (1996) unterscheidet weiter zwischen Performanztests in einem starken und einem schwachen Sinne. Diese Unterscheidung bezieht sich darauf, ob sich der Test ausschließlich auf die sprachliche Leistung bezieht (Performanztest in einem schwachen Sinne) oder ob der Test vor allem die Bewältigung der Aufgabe bewertet (Performanztest in einem starken Sinne). Die Entstehung von Performanztests hängt eng mit der zunehmenden Verbreitung von Fachsprachenprogrammen zusammen. Sprachtests für den Beruf sind häufig Performanztests (z. B. der Test TEACH; siehe Kapitel 5.2, Seite 207 ff). Mit Performanztests vermeidet man die Schwierigkeiten, auf die man beim Einsatz von indirekten Kompetenztests stößt (s. u.). Der Hauptgrund für die Verwendung dürfte in der Validität direkter Performanztests liegen: Durch die Nachbildung der realitätsnahen Sprachverwendungssituation wird eine starke Argumentationsbasis für die Interpretation der Testergebnisse mit Blick auf das Testkonstrukt bereitgestellt. Aus der Beschreibung der sprachlichen Anforderungen, mit der die reale Sprachverwendungssituation zu meistern ist, kann eine mehr oder weniger konkrete Zielvorgabe für den Test abgeleitet werden. Ein weiterer Vorteil liegt in einer möglicherweise hohen Augenscheinvalidität vor allem bei Erwachsenen. Darüber hinaus wird Performanztests eine positive Rückwirkung auf die Lehr- und Lernprozesse zugeschrieben.

Während direkte Performanztests also eine Antwort auf die ersten beiden Fragen von Skehan darstellen könnten, wird die Übertragbarkeit der Ergebnisse auf andere Situatio-

nen nicht gelöst. Dazu gibt ein direkter Performanztest wenig Hinweise. Auch die Reliabilität direkter Performanztests kann ein Problem sein, weil normalerweise die Meinung von Experten oder eine subjektive Beurteilung in die Auswertung einfließt. Einige Autoren stellen auch eine Grundannahme von Performanztests in Frage: Kann ein Sprachtest überhaupt eine reale Sprachverwendungssituation abbilden? Wichtig ist eine möglichst hohe Authentizität, eine Realitätsnähe des Tests. Fehlt diese, ist der Test nicht direkt und die Aussagen über die Validität sind unzutreffend. Auch die Auswahl der Sprachverwendungssituation bzw. der Aufgabe muss repräsentativ für das Testkriterium sein (Baker, 1989; Davies *et al.*, 1999; Grotjahn, 2000a; McNamara, 1996; Messick, 1988; Skehan, 1998).

Sprachliche **Kompetenztests** (*system-referenced test*) haben die allgemeine Beherrschung von Sprache unabhängig von einer konkreten Anwendungssituation zum Gegenstand. In einem Kompetenztest werden Fertigkeiten geprüft, die mit Blick auf verschiedene Konstrukte interpretiert werden können. Kompetenztests liegt eine Vorstellung von Sprache als Code zugrunde, den es zu meistern gilt. Daher sind Kompetenztests häufig als Test isolierter sprachlicher Elemente angelegt. Bei dem Test der Sprachbeherrschung fehlt ein Bezug zu einer konkreten Sprachverwendungssituation. Ausgangspunkt bei der Konstruktion indirekter Kompetenztests ist allein eine Sprachanalyse, ohne Berücksichtigung der Situation (wie bei Performanztests). Man fragt, was es bedeutet, Sprache auf einem bestimmten Niveau zu beherrschen, man fragt, aus welchen Komponenten Sprache besteht und stellt eine repräsentative Auswahl sprachlicher Elemente zusammen. Die Tätigkeiten, die in indirekten Kompetenztests gefordert werden, haben manchmal wenig mit Sprache zu tun: Ankreuzen, einzelne Wörter auf Linien schreiben usw. Doch die häufige Verwendung von geschlossenen Aufgabentypen ermöglicht eine objektive Auswertung und eine hohe Reliabilität. Schwierigkeiten treten bei der Bestimmung der Validität auf. Aufgrund eines eher allgemeinen Testkonstrukts ist die Aussage indirekter Kompetenztests allgemeiner als diejenige von direkten Performanztests. Damit bieten sie eine möglich Antwort auf die dritte Frage von Skehan, sie sind vielseitig einsetzbar. Bei der Interpretation der Ergebnisse verfügt man über einen großen Spielraum. Entscheidungen, die auf der Basis indirekter Kompetenztests getroffen werden, sind jedoch weniger abgesichert, sie können allenfalls auf der Basis von Erfahrung getroffen werden: "In den letzten Jahren haben fremdsprachige Studenten mit dem Ergebnis X im TOEFL große sprachliche Schwierigkeiten im Studium ge-

habt, daher setzen wir nun ein höheres Ergebnis voraus." – Man könnte also erfahrungsbezogen oder mit Verweis auf empirische Untersuchungen argumentieren. Doch die Schwierigkeit liegt wegen ungenauer Aussagen zum abstrakten Testkriterium in der Bestimmung der Konstruktvalidität indirekter Kompetenztests (Baker, 1989; Skehan, 1998).

Die Entscheidung, ob ein Sprachtest als Performanztest oder als Kompetenztest angelegt werden soll, hängt in erster Linie von der Funktion des Tests ab: Soll ein Test nur Auskunft geben über Leistungen der Prüfungskandidaten in einer abgrenzbaren Sprachverwendungssituation, ist ein Performanztest möglich. Wenn die Ergebnisse als Hinweis auf Leistungen in verschiedenen Sprachverwendungssituationen interpretiert werden sollen, wären Kompetenztests eher angemessen. Wohl aus praktischen Erwägungen sind die meisten Sprachtests indirekte Kompetenztests. Dies trifft auch auf Sprachtests für den Hochschulzugang zu, bei denen einzelne Prüfungsteile unterschiedlich ausgelegt sein können.

Die Unterscheidung zwischen **direkten** und **indirekten** Tests hängt in der Praxis eng mit Performanz- und Kompetenztests zusammen, obwohl die Kategorien unterschiedlich sind. Die Klassifizierung direkt vs. indirekt richtet sich allein nach dem Interpretationsbezug. Bei einem direkten Testverfahren kann man ohne weitere Interpretation von der Testleistung auf das Testkriterium schließen, denn die Testleistung und die Leistung in der Wirklichkeit, welche mit dem Testkriterium korrespondiert, entsprechen sich. Je indirekter ein Testverfahren ist, desto größer wird die Notwendigkeit zur Interpretation der Testergebnisse (Bachman, 1990; Davies, 1988: 5-6). Der Unterschied zu dem Begriffspaar Performanz- und Kompetenztest liegt in der Beziehung zum Testkriterium und damit in der unterschiedlichen Beschreibung der Validität: Die Konstruktion von direkten oder indirekten Tests bezieht sich auf die Beziehung des Testinhalts zum Testkriterium. Die Unterscheidung zwischen Performanz- und Kompetenztests beruht in erster Linie auf der unterschiedlichen Vorstellung davon, was es heißt, eine Sprache zu beherrschen. Theoretisch sind Performanztests denkbar, die Anlass für Sprache in Aktion sind, bei denen jedoch in erster Linie nicht geprüft werden soll, ob die Sprachverwendungssituation gemeistert wird, sondern etwas anderes. Dann handelt es sich um einen indirekten Performanztest, bei dem von der Leistung im Test nicht direkt auf das Testkriterium geschlossen werden kann. Die Leistung im Test müsste mit Blick

auf das Testkriterium interpretiert werden. In der Praxis dürften jedoch Performanztests als direkte Tests und Kompetenztests als indirekter Tests entworfen sein.

Die Unterscheidung zwischen direkten und indirekten Tests leidet darunter, dass alle Testverfahren in der Praxis mehr oder weniger indirekt sind. Die Sprachverwendungssituationen in Sprachtests sind künstlich, der Umgang damit erfordert Beurteilungen, Vergleiche und Schlussfolgerungen. Direkte Sprachtests können die Erwartung nicht immer erfüllen, ohne weitere Interpretation auf Leistungen in der Wirklichkeit bzw. in der Zukunft schließen zu können. Fest steht jedoch, dass Sprachtests unterschiedlich indirekt vorgehen, die Nähe zum Testkonstrukt also näher oder weiter sein kann. Die Kritik an der Unterscheidung zwischen direkten und indirekten Sprachtests und an der Erwartung, mit der Verwendung direkter Sprachtests ließe sich die Konstruktvalidität ohne Weiteres in den Griff bekommen, wird übrigens auch zur Unterscheidung zwischen Performanz- und Kompetenztests geäußert, denn die Argumentation zu Performanztests ähnelt der Argumentation zu direkten Tests (Bachman, 1990: 287; Messick, 1996: 244-245; Stevenson, 1985: 116-118).

Sprachtests für den Hochschulzugang sind umfangreiche Testverfahren mit mehreren Prüfungsteilen, die unterschiedlich angelegt sein können. Die Testteile "Textproduktion" und "Mündliche Prüfung" der DSH sind durchaus als Performanztests anzusehen, denn die Aufgaben bilden eine Prüfungssituation im Studium nach. Es handelt sich um Performanztests in einem schwachen Sinne, denn Teilnehmer nehmen mit einer anderen kognitiven Einstellung an einem Sprachtests als an einem Fachtest teil. Die Testteile "Hörverstehen" und "Leseverstehen" sind eher Kompetenztests, weil die Aufgabentypen in der Praxis nicht vorkommen. Beim DSH-Grammatiktest wird nur die sprachliche Richtigkeit bewertet, die Anwendungssituation spielt keine Rolle. Der DSH-Grammatiktest ist ein indirekter Kompetenztest (siehe Kapitel 3.2, Seite 79).

Der TestDaF besteht aus vier Prüfungsteilen, in denen die sprachlichen Fertigkeiten integriert geprüft werden (siehe Tabelle 3). Die Testteile "Schriftlicher Ausdruck" und "Mündlicher Ausdruck" des TestDaF sind sprachliche Performanztests in einem schwachen Sinne. Der kassettengesteuerte Prüfungsteil "Mündlicher Ausdruck" ist ein semidirektes Prüfverfahren, weil die Gesprächsanlässe zwar realitätsnah sind, das Gespräch jedoch von einer Kassette gesteuert wird – keine besonders realistische Kommunikationssituation. Die Testteile "Leseverstehen" und "Hörverstehen" sind eher als Kompe-

tenztests anzusehen. Zwar sind die Lese- und Hörtexte weitgehend authentisch, aber die Aktionen, zu denen die Testteilnehmer angehalten werden, Ankreuzen und Ausfüllen, sind es nicht (Grotjahn, 2000a).

Tabelle 3: TestDaF und DSH – Prüfungsteile im Vergleich

TestDaF	DSH
Leseverstehen Der Prüfungsteil LV besteht aus drei Texten mit unterschiedlichem Schwierigkeitsgrad. Insgesamt sind 30 Items zu bearbeiten (Zuordnung, Mehrfachauswahlaufgaben, Auswahlitems). Die TestDaF-Niveaustufen werden aus den Rohwerten ermittelt (max. 30 Punkte). Dauer: 60 Minuten.	**Verstehen und Bearbeiten eines Lesetextes und wissenschaftssprachliche Strukturen** ("DSH-Leseverstehen" und "DSH-Grammatiktest") Aufgaben zu einem Lesetext, der authentisch, studienbezogen und wissenschaftsorientiert ist. Grammatik ist als textgebundene Aufgabenstellung Teil des Leseverstehens. Die Leistung wird nach sprachlicher Richtigkeit bewertet.
Hörverstehen Der Prüfungsteil HV besteht aus drei Hörtexten mit unterschiedlichem Schwierigkeitsgrad. Insgesamt sind 25 Items zu bearbeiten (gesteuerte Notizen, Auswahlitems). Die TestDaF-Niveaustufen werden aus den Rohwerten ermittelt (max. 25 Punkte). Dauer: 40 Minuten.	**Verstehen und Verarbeiten eines Hörtextes** ("DSH-Hörverstehen") Aufgaben zu einem Hörtext, der max. zweimal präsentiert wird. Der Text soll "der Kommunikationssituation Vorlesung/Übung angemessen Rechnung tragen".
Schriftlicher Ausdruck Beim Prüfungsteil SA bezieht sich die Aufgabenstellung auf eine Grafik. Der Text soll eine Beschreibung und eine begründete Stellungnahme enthalten. Die Bewertung (TestDaF-Niveaustufen) erfolgt durch zwei Prüfer. Dauer: 60 Minuten.	**Vorgabenorientierte Textproduktion** ("DSH-Textproduktion") Texterstellungsaufgabe. Sie kann "erklärender, vergleichender oder kommentierender Art sein, sie kann auch die sprachliche Umsetzung von Grafiken, Schaubildern, Diagrammen zum Gegenstand haben."
Mündlicher Ausdruck Der MA wird als "Simulated Oral Proficiency Interview" (SOPI) durchgeführt. In diesem kassettengesteuerten Prüfungsteil müssen die Kandidaten auf zehn kurze Sprechanlässe reagieren. Bewertung erfolgt in TestDaF-Niveaustufen. Dauer: 30 Minuten.	**Mündliche Prüfung** Prüfungsgespräch, Dauer: max. 20 Minuten. Sie kann entfallen, wenn "für die Beurteilung der mündlichen Kommunikationsfähigkeit hinreichende Erkenntnisse vorliegen."

Quelle: TestDaF: TestDaF-Institut, 2001; Zitate aus der DSH-Rahmenordnung: vgl. FaDaF, 2001.)

Modelle von Sprachkompetenz

Sprachtests werden eingesetzt, um ein Bild über die Sprachkompetenz von Kandidaten zu gewinnen. Auf dieser Basis werden Entscheidungen getroffen. Testanwender müssen entscheiden, wie die Leistungen im Test mit Leistungen in anderen Sprachverwendungssituationen korrespondieren. Bei der Interpretation der Ergebnisse von Sprachtests muss man daher auf Modelle zurückgreifen, welche den Zusammenhang zwischen der Leistung im Test und der Leistung außerhalb des Tests verdeutlichen. Diese Notwendigkeit besteht vor allem bei indirekten Kompetenztests, bei denen der Interpretationsspielraum groß ist. Diese werden – bewusst oder unbewusst – mit Blick auf ein Modell von Sprachkompetenz konstruiert und eingesetzt. Wie im Abschnitt zu den Performanztests bereits angesprochen wurde, beruhen auch direkte Performanztests auf einem Modell der Sprachkompetenz. Doch die Bezugnahme auf das Modell steht nicht im Mittelpunkt.

Es gibt miteinander konkurrierende Modelle, welche sich widersprechen bzw. andere Schwerpunkte setzen. Das Sprachverständnis von Lado (1961; 1971) war von der amerikanischen Linguistik geprägt. Der strukturelle Deskriptivismus verstand Sprache als hierarchisches System von mehr oder weniger isolierten Elementen. Er erläutert:

> The matter to be tested is language. Language is built of sounds, intonation, stress, morphemes, words, and arrangements of words having meanings that are linguistic and cultural. The degree of mastery of these elements does not advance evenly but goes faster in some and slower in others. Each of these elements of language constitutes a variable that we will want to test. They are pronounciation, grammatical structure, the lexicon, and cultural meanings (Lado, 1961: 25).

Die Konzeption der Sprachtests wurde außerdem durch die psychometrische Testtheorie beeinflusst, welche sich unabhängiger (gesonderter) Items (*discrete point items*) bediente. Es entstanden indirekte Kompetenztests, die auf dem Modell beruhten, dass sich die sprachlichen Elemente wie auch die sprachlichen Fertigkeiten mehr oder weniger isoliert voneinander entwickeln können und daher separat zu prüfen sind. Die Tests bestanden demnach aus einer möglichst umfassenden Auswahl verschiedener sprachlicher Elemente. Das theoretische Fundament dieses Ansatzes überzeugte zunächst, auch die hohe Objektivität und Reliabilität der verwendeten Sprachtests waren nicht zu leugnende Vorzüge. Unzufriedenheit entstand vor allem aufgrund der verwendeten Testformate und der fragwürdigen Validität der Testergebnisse (Baker, 1989; Bolton, 1997; Lado, 1961).

Aus der Kritik am strukturalistisch-psychometrischen Sprachverständnis, welche sich vor allem gegen das Testformat richtete, entstand Ollers These von der Eindimensionalität der Sprachkompetenz (*Unitary Competence hypothesis*). Dabei bezog er sich auf Modelle von Chomsky und Spolsky. Das linguistische Modell von Chomsky (1964; 1965; 1972) beruht auf folgender Vorstellung: Der Sprache liegen Regeln zugrunde, welche von den Sprechern beherrscht werden. Sprachliche Leistungen und die Beherrschung der Regeln stehen daher in einem Verhältnis. Ähnlich äußerte sich auch Spolsky:

> Knowing a language is a matter of having mastered (as yet incompletely specified) rules; the ability to handle new sentences is evidence of knowing the rules that are needed to generate them (Spolsky, 1973: 173).

Während die Sprachkompetenz in der Vorstellung von Chomsky und Spolsky durchaus mehrdimensionalen Charakter haben konnte, ging Oller einen Schritt weiter und behauptete, dass die Sprachkompetenz auf einen gemeinsamen Faktor zurückgehe. Unterschiedliche Ergebnisse von Sprachtests führte er darauf zurück, dass die Messungen dieses gemeinsamen Faktors jeweils unterschiedlich effektiv seien. Er fühlte sich durch die Ergebnisse von Faktoranalysen bestätigt (Erläuterung von "Faktoranalysen" siehe Kapitel 4.2.1, Seite 125). Führt man Faktoranalysen mit mehreren sprachlichen Variablen durch, so kann man häufig nur einen Faktor mit einem hohen Eigenwert extrahieren, auf den die einzelnen Variablen mehr oder weniger stark laden. Ein zentrales Konzept seines Modells war die "pragmatische, erwartbare Grammatik" (*pragmatic expectancy grammar*): "In the meaningful use of language, some sort of pragmatic expectancy grammar must function in all cases" (Oller, 1979: 25). Diese Fähigkeit ermöglicht laut Oller die Sprachproduktion und das Sprachverständnis. Aufgabe von Sprachtests war es dann, diese zentrale Sprachfähigkeit abzubilden.

Eine spätere Analyse der Faktoranalysen von Oller führte jedoch zu weniger eindeutigen Ergebnissen, außerdem wurden andere Erklärungsmodelle entwickelt (siehe unten). Die These von der eindimensionalen Sprachkompetenz wurde von Oller selbst widerrufen ("The idea of an exhaustive global factor of language proficiency was wrong"; Oller, 1983a: 35). Sie überzeuge nicht, weil sie nicht mit den praktischen Erfahrungen übereinstimmte. Weder die Vorstellung einer völlig unabhängigen Entwicklung der Kompetenz in den vier Fertigkeiten noch die Vorstellung einer völligen Abhängigkeit voneinander entspricht der Erfahrung von Fremdsprachenlernern und –lehrern (Baker,

1989; Oller, 1974; 1976; 1979; 1981; 1983a; 1983b; Sang/Vollmer, 1978; Vollmer, 1983).

Das Modell von Oller spielt in der wissenschaftlichen Diskussion kaum noch eine Rolle. Verwirrung entsteht bisweilen beim Einsatz von "integrativen" Tests, welche in der Praxis so eingesetzt werden, als träfe die These der Eindimensionalität der Sprachkompetenz zu. Carroll (1961) entwickelte bereits in den 1960er Jahren integrative Tests. Diese bestanden aus einer mündlichen Prüfung, einer Textproduktion und einem Diktat. Auch der C-Test (Erläuterung siehe Kapitel 1, Seite 8) muss manchmal als Test der eindimensionalen Sprachkompetenz herhalten. Beim C-Test sind hohe Korrelationen mit Tests unterschiedlicher Fertigkeiten zu beobachten und C-Tests laden bei Faktoranalysen hoch auf den ersten Faktor, der häufig als "allgemeine Sprachkompetenz" interpretiert wird. Doch dies ist der Komplexität des C-Tests zuzuschreiben, bei dem sehr unterschiedliche Kompetenzen zum Einsatz kommen, und nicht der Eindimensionalität der Sprachkompetenz (Fotos, 1991; Grotjahn, 1992; 1994; 1995; 2002; Coleman/Grotjahn/Raatz, 2002).

Mit dem Ende der Hypothese von der Eindimensionalität der Sprachkompetenz (*Unitary Competence Hypothesis*) ist die Diskussion um die Struktur der Sprachkompetenz nicht beendet. Die folgenden Modelle der kommunikativen Sprachkompetenz haben gemein, dass sie von einer Mehrdimensionalität der Sprachkompetenz ausgehen, bei der einzelne Komponenten mehr oder weniger stark in Beziehung stehen. Es ist das Ziel dieser Modelle, den Zusammenhang zwischen sprachlicher Leistung und zugrunde liegenden Fähigkeiten zu erläutern.

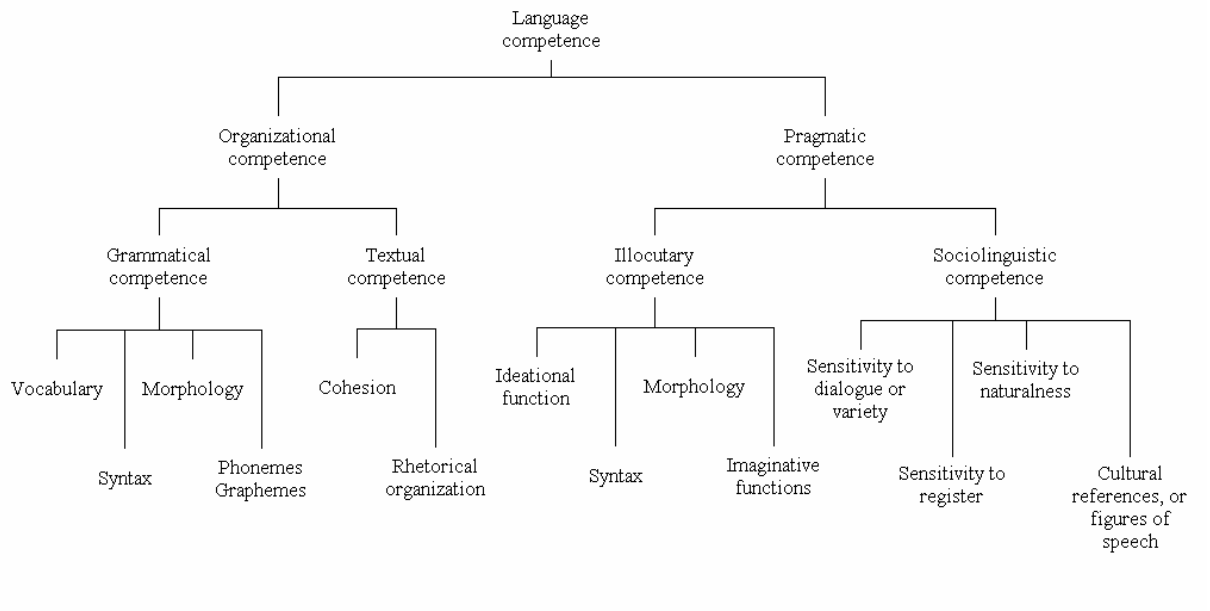
Widdowson (1978; 1979; 1983) bzw. Canale und Swain (1980) beschreiben Sprachkompetenz als mehrdimensional und dynamisch, und sie beziehen diskursive und soziolinguistische Aspekte mit ein. Canale (1983) zählt vier Komponenten der Sprachkompetenz auf: linguistische, soziolinguistische, diskursive und strategische Kompetenzen. Unter der linguistischen Kompetenz versteht er analog zu Chomsky die Beherrschung des Regelsystems, das der Sprache zugrunde liegt. Unter der soziolinguistischen Kompetenz versteht er den situationsangemessenen Gebrauch von Sprache. Die diskursive Kompetenz bezieht sich auf die Fähigkeit, Sprache auch über den Einzelsatz hinaus einzusetzen, also auf den Umgang mit Texten und komplexen Äußerungen. Die strategische Kompetenz hat nach der Auffassung von Canale und Swain hauptsächlich

kompensatorischen Charakter. Sie kommt vor allem dann zum Einsatz, wenn die anderen Kompetenzen nicht ausreichen (Canale, 1983; Canale/Swain, 1980).

Das Modell von Bachman (1990) bzw. Bachman und Palmer (1996) ist als eine direkte Weiterentwicklung und Ausdifferenzierung der Vorstellungen von Canale und Swain zu sehen. Eine Besonderheit ist der ausdrückliche Bezug zu Sprachtests. Das Modell weist im Unterschied zu demjenigen von Canale und Swain folgende Besonderheiten auf: Bachman und Palmer unternehmen eine sehr komplexe Beschreibung der Sprachkompetenz und seiner Komponenten. Sie weisen außerdem der strategischen Kompetenz eine zentrale Rolle zu. Die Sprachkompetenz besteht laut Bachman aus folgenden Komponenten: Die organisatorische Kompetenz besteht aus der grammatischen und der textuellen Kompetenz, die pragmatische Kompetenz besteht aus der illokutiven und soziolinguistischen Kompetenz (siehe Abbildung 2). Im Vergleich zu vorhergehenden Modellen wird hier ein differenzierteres Bild der Komponenten von Sprachkompetenz entworfen, was vor allem die pragmatische Kompetenz betrifft.

Sprachfähigkeit beruht laut Bachman und Palmer nicht nur auf Sprachkompetenz, sondern auch auf strategischer Kompetenz. Bachman (1990) spricht von *language competence* – Sprachkompetenz, Bachman und Palmer (1996) sprechen von *language knowledge* – Sprachwissen. Der strategischen Kompetenz wird bei Bachman und Palmer nicht nur ein kompensatorischer Charakter zugeschrieben, sie stellt vielmehr ein zentrales Bindeglied für jede Form der Kommunikation dar. Sie ermöglicht sprachliche Äußerungen im Spannungsfeld von Sprachkompetenz (oder Sprachwissen), Sachkenntnissen und der jeweiligen Kommunikationssituation. Insgesamt betonen Bachman und Palmer die Wechselwirkung zwischen den Hauptkomponenten der Sprachfähigkeit, so dass die Beziehung zwischen sprachlicher Leistung und sprachlichem Wissen als eine dynamische angesehen wird.

Das Verdienst dieses Modells liegt darin, eine Übersicht über mögliche Komponenten der Sprachkompetenz zu bieten, die als weitgehend empirisch abgesichert gelten und bei der Testerstellung und Testanalyse als Instrument zur Vergewisserung der Vorgehensweise dienen kann.



Quelle: Bachman, 1990: 87.

Abbildung 2: Bestandteile der Sprachkompetenz nach Bachman (1990)

Skehan kritisiert an dem Modell von Bachman und Palmer, dass es den bloßen Charakter einer Liste zum Abhaken annehme und eine Gewichtung der Komponenten noch ausstehe:

> The current status of the framework seems to be that it proposes categories which function as a systematizing checklist, but which do not give clear indications of significance, centrality, or relative importance (Skehan, 1998: 163-164).

Schließlich vermisst er eine Einbettung in psycholinguistische Mechanismen, die er in einem eigenen Modell der Sprachverarbeitung (*processing approach*) zentral berücksichtigt. Er bezieht sich dabei auf ein psycholinguistisches Konstrukt, das den Unterschied zwischen sprachlichen Fähigkeiten und sprachlicher Leistung erklären kann, nämlich die Fähigkeiten zum Einsatz von Sprache (*abilities for use*). Bei Sprachtests, welche dem Modell der Sprachverarbeitung folgen, geht es dann nicht um die möglichst repräsentative Auswahl von sprachlichen Fähigkeiten, welche der sprachlichen Leistung zugrunde liegen, sondern um die Auswahl von Faktoren, welche einen Einfluss auf die

sprachliche Leistung haben. Das Ergebnis dieser Analyse sind auf jeden Fall direkte Performanztests, wobei sich Skehan ausdrücklich auf die schwache Form der Performanztests bezieht.

Modelle von Sprachkompetenz können auf eine konkrete Sprachverwendungssituation bezogen werden; bei Sprachtests für den Hochschulzugang geht es um Sprachkompetenz im Fachstudium. Betrachtet man etwa das Modell von Bachman (bzw. Bachman und Palmer), so ist zu fragen, welchen Aspekten eine besondere Bedeutung zukommt. Die beiden Aspekte, die im Mittelpunkt der vorliegenden Arbeit stehen, Grammatik und Fachkenntnisse, gehören meiner Ansicht nach zur Sprachkompetenz im Studium. Dies erklärt jedoch noch nicht, ob und in welcher Weise beide Aspekte in Sprachtests für den Hochschulzugang berücksichtigt werden sollten. Nach dem Modell der Sprachverarbeitung von Skehan würde man fragen, in welchem Umfang Grammatik und Fachkenntnisse einen Einfluss auf die sprachliche Leistung im Studium haben. Es muss weiter geklärt werden, ob andere Faktoren den Einfluss der Grammatik (bzw. der Fachkompetenz) bereits erfassen. Dies sind Fragestellungen, denen ich in dieser Arbeit nachgehe. Zunächst ist festzuhalten, dass den beiden Aspekten in der DSH und im TestDaF durchaus unterschiedliche Bedeutungen beigemessen werden.

Zusammenfassung

Mit dem TestDaF und der DSH stehen ausländischen Studienbewerbern zwei unterschiedliche Sprachtests für den Nachweis ausreichender Deutschkenntnisse zur Verfügung. Die DSH ist im Gegensatz zum TestDaF nicht standardisiert. Sie hat den Charakter einer Zulassungsprüfung. In der Praxis dürfte sie häufig auch Züge einer Kursabschlussprüfung annehmen, was ungewöhnlich und eigentlich nicht angebracht ist. Der TestDaF ist eine Feststellungsprüfung, die freilich von den Abnehmern wie eine Zulassungsprüfung verwendet wird. Möglicherweise bietet der TestDaF durch den differenzierten Ergebnisausweis in dieser Hinsicht mehr Informationen als notwendig sind. DSH und TestDaF sind kriteriumsbezogene Prüfungen, bei denen die einzelnen Prüfungsteile unterschiedlich direkt oder indirekt ausgelegt sind. Neben der fehlenden Standardisierung der DSH ist die Existenz von Items zur Grammatik ein Unterschied zum TestDaF.

2.2. Nützlichkeit von Sprachtests für den Hochschulzugang

Übersicht: Kapitel 2.2
Auf die Einordnung der DSH und des TestDaF nach testtheoretischen Aspekten folgt eine kritische Analyse beider Sprachtests anhand der Kriterien zur Nützlichkeit von Sprachtests von Bachman und Palmer. Dazu zählen Reliabilität, Konstruktvalidität, Authentizität, Interaktivität, Testauswirkungen und Testökonomie.

Der Kriterienkatalog von Bachman und Palmer (1996; siehe Tabelle 4, Seite 35 ff) ist eine Weiterentwicklung der klassischen Testtheorie für Sprachtests. Die klassische Testtheorie beruht auf bestimmten Annahmen über gemessene und wahre Testwerte sowie über die Faktoren, welche die gemessenen Testwerte beeinflussen (Messfehler). Ausgehend von Annahmen über Eigenschaften des Messfehlers wird die Qualität eines Tests durch die Hauptkriterien Objektivität, Reliabilität und Validität bestimmt (Bachman, 1990; Bortz/Döring, 2002; Ingenkamp, 1985; Lienert/Raatz, 1994). Die Gütekriterien der klassischen Testtheorie werden häufig als nicht umfassend genug angesehen. Die Hauptkriterien sind von einigen Autoren bereits differenziert und erweitert worden. Lienert und Raatz führen beispielsweise Normierung, Vergleichbarkeit, Ökonomie und Nützlichkeit von Sprachtests als "Nebengütekriterien" an (1994: 7-14). Die klassische Testtheorie konzentriert sich jedoch nicht in erster Linie auf Sprachtests, sondern allgemein auf psychometrische Tests – zu denen man nicht nur Sprachtests, sondern auch andere Tests für psychologische Merkmale wie Persönlichkeitsmerkmale oder Intelligenz zählt. Die Konzeption von Bachman und Palmer, welche ausdrücklich mit Blick auf Sprachtests konzipiert wurde, stimmt grundsätzlich durchaus mit der klassischen Testtheorie überein, setzt aber andere Schwerpunkte. Als theoretische Grundlage beziehe ich mich bei meiner Analyse von Sprachtests für den Hochschulzugang auf die von Bachman und Palmer vorgestellten Kriterien.

Zur Qualität eines Sprachtests tragen laut Bachman und Palmer folgende Komponenten bei: Reliabilität, Konstruktvalidität, Authentizität, Interaktivität, Auswirkungen und Ökonomie ("*reliability, construct validity, authenticity, interactiveness, impact, and practicality*", 1996: 9). Allerdings sprechen Bachman und Palmer nicht von der "Qualität" eines Sprachtests, sondern von *usefulness* ("Nützlichkeit" oder "Brauchbarkeit"). Bachman und Palmer regen an, die einzelnen Komponenten nicht isoliert, sondern nur unter Berücksichtigung der Wechselwirkungen und der Auswirkungen auf die Nützlichkeit auszuwerten. Die einzelnen Komponenten sollten in einem ausgewogenen Verhältnis stehen (Bachman/Palmer, 1996: 17-42).

Zur Übersetzung ist anzumerken, dass *usefulness* in der deutschen Literatur mit "Nützlichkeit" oder "Brauchbarkeit" wiedergegeben wird (Grotjahn, 2000a: 320). Das ist zwar korrekt, doch geht durch diese Übersetzung meiner Ansicht nach ein Teil der Begründung für die Wahl dieses Begriffs verloren:

> The most important consideration in designing and developing a language test is the use for which it is intended, so that the most important quality of a test is its usefulness. This may seem so obvious that it need not be stated (Bachman/Palmer, 1996: 17).

Die Beziehung zwischen "*intended use*" und "*usefulness*" lässt sich im Deutschen eher mit den Begriffen "Funktion" und "Funktionalität" oder "Zweck" und "Zweckmäßigkeit" wiedergeben. Ich verwende aber den Begriff "Nützlichkeit", da er sich in der deutschsprachigen Literatur durchgesetzt hat.

Tabelle 4: Kriterien für die Nützlichkeit von Sprachtests nach Bachman/Palmer

Kriterium	Beschreibung
Reliability Reliabilität	*The consistency of measurement* Die Konsistenz oder Stabilität der Messungen
Construct validity Konstruktvalidität	*The appropriateness of the interpretations made on the basis of test scores* Die Gültigkeit der auf der Basis der Testergebnisse vorgenommenen Interpretationen
Authenticity Authentizität	*The degree of correspondence of the characteristics of a given language test task to the characteristics of a target language use task* Ausmaß, in dem der Test Sprachbenutzung außerhalb einer Testsituation widerspiegelt
Interactiveness Interaktivität	*The extent of involvement of the test taker's language ability, topical knowledge, and interest in accomplishing a test task* Ausmaß, in dem die Sprachfähigkeit der Kandidaten, Wissen über das Thema und das Interesse an der Erfüllung der Aufgabe einbezogen werden
Impact Auswirkungen	*Ways in which test use affects society, an education system, and the individuals within these* Auswirkungen einer Prüfung auf die Gesellschaft, das Bildungssystem und Einzelpersonen
Practicality Ökonomie	*Relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities* Verhältnis zwischen den Mitteln, die für die Konzeption, Entwicklung und den Einsatz des Tests benötigt werden und den Mitteln, die dafür zur Verfügung stehen

(nach Bachman/Palmer, 1996: 38-40)

Reliability – Reliabilität

Die Reliabilität bezieht sich auf die Konsistenz und Stabilität, mit der ein Merkmal gemessen wird. Auch wenn an anderer Stelle die Rolle der Testfunktion betont wurde: Für die Reliabilität ist sie unbedeutend. Die hohe Reliabilität eines Tests gibt lediglich Auskunft darüber, dass er in Zukunft die gleichen Ergebnisse liefert wie in der Vergangenheit, sie sagt nichts darüber aus, ob die Ergebnisse auch relevant sind. Dennoch darf die Reliabilität nicht als Kriterium für die Qualität von Sprachtests fehlen, denn sie ist eine Voraussetzung für einen funktionalen Einsatz, also auch für die Validität. In Abhandlungen zur Nützlichkeit (bzw. Qualität) von Sprachtests wird ausnahmslos darauf hingewiesen, dass Ergebnisse, die nicht zuverlässig sind, keine Schlussfolgerung gemäß der Messintention zulassen. Als Beispiel sei Bachman angeführt:

> If test scores are strongly affected by errors of measurement, they will not be meaningful, and cannot, therefore, provide the basis for valid interpretation or use. A test score that is not reliable, therefore, cannot be valid (1990: 25).

Man unterscheidet in der klassischen Testtheorie verschiedene Aspekte der Reliabilität: Die interne Konsistenz (*internal consistency*) bezieht sich auf Unterschiede zwischen den einzelnen Items. Inkonsistente Items würden beispielsweise den Testteilnehmern völlig unterschiedliche Rangplätze zuweisen. Die Testwiederholungsreliabilität (*stability*; *test-retest reliability*) bezieht sich z. B. auf die Vergleichbarkeit der Testergebnisse, wenn der Test nach Ablauf einer gewissen Zeit wiederholt wird. Die Paralleltestreliabilität (*equivalence; parallel forms reliability*) bezieht sich auf die Vergleichbarkeit von verschiedenen Versionen eines Tests (Bachman, 1990: 160-186).

Eine Bestimmung der Reliabilität ist empirisch möglich. Um Informationen zur internen Konsistenz zu gewinnen, werden nach der Methode der Testhalbierung die Testergebnisse der beiden Hälften korreliert (Split-Half-Reliabilität) oder einzelne Items mit anderen korreliert. Statistisch wird die Reliabilität beispielsweise über Cronbachs alpha-Koeffizienten erhoben. Die Testwiederholungsreliabilität kann über Korrelationen zwischen den Ergebnissen bestimmt werden, welche Testeilnehmer in demselben Test in zwei Durchläufen erzielten. Außerdem können die Mittelwerte und die Standardabweichungen verglichen werden. Auch die Paralleltestreliabilität wird über Korrelationen ermittelt. Dabei werden die Ergebnisse verglichen, welche die Testteilnehmer in beiden Tests erzielen. Diese sollten zeitnah und möglichst auch in einer Kontrollgruppe in unterschiedlicher Reihenfolge durchgeführt werden.

Welche Aspekte der Reliabilität sind für Sprachtests für den Hochschulzugang besonders wichtig? Bei umfangreichen Prüfungen wie Sprachtests für den Hochschulzugang dürfte eine Testwiederholungsreliabilität nie bei 100 Prozent liegen, denn Menschen sind keine Maschinen, die eine absolut konstante Leistung bringen. Die Leistungsschwankungen durch unterschiedliche Stimmungen oder unterschiedliche Motivation führt zu einer Reliabilität unter 100 Prozent. Auch die interne Konsistenz dürfte nicht sehr hoch sein. Dies ist jedoch nicht unbedingt ein Zeichen für mangelnde Qualität, es ist vielmehr eine Folge des vielschichtigen Testkonstrukts, welches zu einem eher heterogenen Test führt. Die interne Konsistenz hängt von der Homogenität eines Tests ab. Bei homogenen Tests wird ein einziges Merkmal mit verschiedenen Items gemessen, bei heterogenen Tests werden unterschiedliche Merkmale mit verschiedenen Items gemessen. Die interne Konsistenz ist bei heterogenen Tests zwangsläufig niedrig. Geht man davon aus, dass die Sprachkompetenz vielgestaltig ist,

so handelt es sich bei Sprachtests für den Hochschulzugang um heterogene Tests. Allein bei einzelnen Prüfungsteilen ist eine hohe Reliabilität zu erwarten (Bortz/Döring, 2002: 195-199; Brown, 2001: 173-176; Lienert/Raatz, 1994: 137-219). Beide Aspekte der Reliabilität, die Testwiederholungsreliabilität und die interne Konsistenz, haben bei Sprachtests für den Hochschulzugang eine Bedeutung. Die Paralleltestreliabilität ist für die DSH ein besonders wichtiges Thema. Für die Testteilnehmer ist es von großer Bedeutung, ob die Version, die ihnen zur Bearbeitung vorgelegt wird, zu dem gleichen Ergebnis kommt, wie eine andere Version.

Die geringe Paralleltestreliabilität ist potenziell die Achillesferse von dezentralen, nicht standardisierten Prüfungen wie der DSH. Möglicherweise sind einzelne Prüfungsversionen *innerhalb einer Institution* noch relativ reliabel, da sie von denselben Personen erstellt und bewertet werden. Wenn die Prüfer gleich bleibende Bewertungsmaßstäbe anlegen und die Prüfungen mit einer gewissen Konsistenz erstellen, dürfte sich die Messgenauigkeit nicht allzu sehr unterscheiden. Ob es gelingt, bei dezentralen Prüfungen *zweier Institutionen* eine hohe Reliabilität zu erreichen, darf bezweifelt werden, selbst wenn diese nach einem bestimmten Muster entworfen wurden. Ich gehe davon aus, dass die Paralleltestreliabilität zwischen DSH-Prüfungen nicht gesichert ist, wenn sie von unterschiedlichen Institutionen entworfen und korrigiert werden. Dies wurde in einer Studie zum Hörverstehen eindrucksvoll bestätigt. Koreik und Schimmel (2002) ließen 24 ausländische Studienbewerber an fünf verschiedenen DSH-Hörverstehenstests teilnehmen. Nur bei zehn Testteilnehmern waren die Ergebnisse eindeutig: Sie fielen in allen fünf Tests durch. Alle anderen Testteilnehmer bestanden einige Tests und fielen in anderen durch. Das Hörverstehen ist eine komplexe Fertigkeit und die Leistungen in dieser Fertigkeit können durchaus Schwankungen unterworfen sein. Allerdings stellten Koreik und Schimmel bei den Prüfungsteilen zum Hörverstehen bedeutende Unterschiede fest, die fraglos Auswirkungen auf die Paralleltestreliabilität haben: In einer DSH wurde der Hörtext einmal vorgetragen, in drei DSH-Prüfungen zweimal und in einer gar dreimal. Hörtexte (Länge, Anspruchsniveau), Items und Auswertungskategorien wiesen ebenfalls deutliche Unterschiede auf.

Zur Reliabilität der DSH sollen außerdem einige Faktoren diskutiert werden, welche die Reliabilität von Tests beeinflussen (Hughes, 2003; 44-50: Lienert/Raatz, 1994: 202-213):

- Die Reliabilität von Tests hängt von der Objektivität der Auswertung ab. Dazu wäre es nötig, die Prüfer zu trainieren, ihnen einheitliche Auswertungsmaßstäbe an die Hand zu geben und ihre Arbeit zu evaluieren. Mir sind Fortbildungsangebote für DSH-Korrekteure nicht bekannt, sie finden allenfalls informell unter den Prüfern statt. Alle TestDaF-Korrekteure der Prüfungsteile Sprechen und Schreiben werden demgegenüber in Seminaren auf ihre Aufgaben vorbereitet. Anhand von Probekorrekturen wird außerdem die Prüferstrenge bzw. –milde eruiert, welche bei der Ergebniserstellung berücksichtigt wird (Eckes, 2003a; 2003b). Ein weiterer Gesichtspunkt muss beachtet werden: Wenn die Prüfer Informationen über die Testteilnehmer haben oder die Kandidaten kennen, kann ihre Beurteilung durch bestimmte Erwartungen an die Leistungen der Kandidaten beeinträchtigt werden. Sowohl beim TestDaF als auch bei der DSH sehen die Korrekteure den Namen der Testteilnehmer. Denkbar ist, dass sie mit bestimmten Nationalitäten bestimmte Leistungen verbinden: "Kandidaten aus XY können gar nicht so gut/so schlecht sein!" Verstärkt wird das Problem bei der DSH, da die Kandidaten den Prüfern häufig aus dem Unterricht bekannt sind.
  Bachman und Palmer (1996: 19-21) führen "Objektivität" übrigens nicht gesondert als Testgütekriterium an, weil sie die Objektivität der Testdurchführung zur Reliabilität zählen (*rater consistency*). Mit der Objektivität wird in der klassischen Testtheorie die Unabhängigkeit vom Testanwender bezeichnet (Bortz/Döring, 2002: 194-195; Lienert/Raatz, 1994: 7). Wenn Diskrepanzen zwischen Testanwendern auftreten, wird die Reliabilität des Tests beeinträchtigt, denn die Messergebnisse werden ungenauer, was zum Beispiel die Testwiederholungsreliabilität beeinträchtigen würde. Eine Abhängigkeit vom Testanwender wird von Bachman und Palmer als eine von vielen Störgrößen angesehen, welche die Reliabilität schmälern können. Die Reliabilität kann nur so hoch sein, wie die Objektivität, einem ihrer konstitutiven Elemente (Bachman, 1990: 172-178; Hughes, 2003; 44-50: Lienert/Raatz, 1994: 202-213).

- Die Reliabilität eines Tests hat mit der Streuung der Aufgabenschwierigkeit zu tun. Die Reliabilität wird durch Items beeinträchtigt, die nicht stark zwischen Testteilnehmern mit starken und schwachen Leistungen differenzieren. Dieses Problem lässt sich auf zwei Weisen in den Griff bekommen: Erstens durch eine Testverlängerung und zweitens durch eine Identifizierung und Änderung wenig differenzierender

Items im Zuge von Erprobungsfassungen. Die Durchführung von Erprobungen ist bei der DSH jedoch eher die Ausnahme, beim TestDaF werden vor dem Einsatz eines neuen Tests hingegen umfangreiche Erprobungen durchgeführt.

- Die fehlende Vertrautheit der Testteilnehmer mit dem Format einer Prüfung ist eine weitere Ursache für eine Minderung der Reliabilität. Ob die Aufgabenstellungen von den Testteilnehmern verstanden werden, lässt sich durch Erprobungsfassungen feststellen. Wenn die Kandidaten in institutseigenen Kursen auf die DSH vorbereitet werden, ist eine Vertrautheit mit dem Format zu erwarten. Bei externen Testteilnehmern stellt sich die Situation anders dar. Sie können sich bei der dezentralen DSH nicht darauf verlassen, dass das erwartete Format auch verwendet wird. Sie sind auf Musterprüfungen angewiesen, die mittlerweile von fast allen Hochschulen angeboten werden, welche die DSH durchführen. Hier kann es zu einer Beeinträchtigung des Ergebnisses durch eine unverständliche Aufgabenstellung oder die fehlende Vertrautheit mit dem Format kommen. Verwirrende Aufgabenstellungen dürften beim TestDaF nach den Erprobungen nicht mehr auftreten. Es kann zwar vorkommen, dass sich die Kandidaten vorher nicht mit dem Format vertraut gemacht haben. Wenn sie sich aber mit dem Format auseinander gesetzt haben, können sie sich darauf verlassen, dieses Format auch anzutreffen. Unterschiede zwischen den einzelnen TestDaF-Prüfungszentren gibt es nicht (Eckes, 2003a).

Perlmann-Balme weist auf die innerinstitutionelle Funktion der DSH hin, welche dazu führe, dass die Reliabilität kein zentrales Argument für die Nützlichkeit darstelle:

> Die Frage von Abweichungen im Schwierigkeitsgrad bei abweichenden Prüfungsformaten von Universität zu Universität stellt sich insofern nicht als Problem, als jede Universität zugleich Prüfungsmacher und Endabnehmer der Zeugnisse, also zugleich prüfende und anerkennende Institution ist (Perlmann-Balme, 2001: 1001).

Da die DSH bundesweite Anerkennung hat und da sie von den Studienbewerbern auch zu bundesweiten Bewerbungen um einen Studienplatz verwendet wird, dürften immer auch Studierende an den Hochschulen sein, welche nicht die hochschuleigene DSH absolviert haben. Darüber hinaus gibt es eine Reihe von Hochschulen, welche keine eigene DSH abnehmen. Insofern ist die fehlende Reliabilität der DSH als gravierender Mangel anzusehen.

Die geringe Paralleltestreliabilität der DSH wird von den Kandidaten erkannt und führt dazu, dass sie sich häufig an mehreren Standorten zur DSH anmelden. Dieser

"Prüfungstourismus" wäre überflüssig, wenn die Reliabilität der DSH hoch wäre. In den letzten Jahren gab es vom Fachverband Deutsch als Fremdsprache (FaDaF) mehrere Initiativen, welche das Ziel hatten, die Reliabilität der DSH zu erhöhen. Um einheitliche Testmethoden-Merkmale zu etablieren, wird die DSH-Rahmenordnung seit 2001 durch detaillierte Ausführungen im DSH-Handbuch erweitert. Seit 2005 müssen DSH-Ausrichter die örtliche Prüfungsordnung und das Zeugnis beim FaDaF registrieren lassen, wenn sie zu den anerkannten DSH-Ausrichtern zählen wollen. Dass diese Maßnahmen nicht ausreichen, die Reliabilität der DSH grundlegend zu verändern, ist allen Beteiligten bewusst: Ziel dieser Maßnahmen ist lediglich, die Unterschiede zwischen den DSH-Prüfungen zu reduzieren. Weitere Maßnahmen zur Verbesserung der Reliabilität wie etwa die Überprüfung der Prüfungen auf einheitliche Testmethoden-Merkmale oder die Überprüfung von Beispielarbeiten durch eine Zentralstelle wären aus Sicht der Testtheorie weitere notwendige Schritte, die höchstwahrscheinlich von DSH-Ausrichtern nicht akzeptiert würden.

Wenn man die DSH aus der Perspektive der Testtheorie betrachtet, muss die bundesweite Anerkennung in Frage gestellt werden. Da man die Reliabilität der DSH nicht sicherstellen kann, ist die Forderung nach der Anerkennung der DSH an allen Hochschulen in Deutschland nicht gerechtfertigt.

Diese geringe Reliabilität der DSH war ein Grund für die Entwicklung des TestDaF (Projektgruppe TestDaF, 2000). Beim TestDaF sind mehrere Maßnahmen vorgesehen, um eine hohe Reliabilität zu sichern (Arras/Grotjahn, 2002). Eine Maßnahme sind Probedurchläufe. Die Teilnehmer an den Probedurchläufen bearbeiten nicht nur die neue Prüfungsversion, sondern auch einen immer gleichen C-Test. Die Ergebnisse aus diesem C-Test werden als Basis für die Einschätzung des jeweiligen Schwierigkeitsgrads eingesetzt. Das Verfahren stellt – abgesehen von unvermeidbaren Messfehlern – eine relativ hohe Reliabilität her. Im Übrigen orientiert sich der TestDaF bei der Bestimmung der Reliabilität ähnlich wie andere große Sprachtests an der probabilistischen Testtheorie. Die probabilistische Testtheorie ist eine Weiterentwicklung des klassischen Modells. Sie beruht auf Annahmen über die Wahrscheinlichkeit, mit der bestimmte Antworten auftreten. Faktoren sind dabei nicht nur die Itemschwierigkeit, sondern auch die individuellen Fähigkeiten der Kandidaten. Mit Hilfe von mathematischen Modellen über die Antwortwahrscheinlichkeit lassen sich mehrere Ursachen

für Varianzen gleichzeitig beobachten und Informationen zur Reliabilität kombiniert zusammentragen (Bachman, 1990: 187-227; Davidson, 2000; Fischer/Molenaar, 1995; Fisseni, 1990: 116-142; Pollitt, 1997; Shavelson/Webb, 1991).

Construct validity – Konstruktvalidität

Unter Validität versteht man das Ausmaß, in dem der Test Schlussfolgerungen gemäß der Messintention zulässt. Ein Test ist also nicht grundsätzlich valide, die Validität kann nur mit Blick auf eine bestimmte Funktion bestimmt werden. Über das Verständnis von Validität eines Tests, die unbestreitbar ein konstitutives Kriterium für die Nützlichkeit von Sprachtests darstellt, gibt es unterschiedliche Vorstellungen. Der Kern von Lados Definition (s. u.) hat zwar noch Gültigkeit, die Aufmerksamkeit richtet sich inzwischen jedoch auf weitere Aspekte. Für Lado ergibt sich die Validität aus der Relevanz:

> Does the test measure what it claims to measure? If it does, it is valid. [...] In other words, for a test to be valid we expect the content and conditions to be relevant, and that there will be no irrelevant problems which are more difficult than the problems being tested (Lado, 1961: 321).

Das derzeitige Verständnis von Validität hat sich im Vergleich zu Lados Definition in einigen Punkten geändert. Die Beiträge von Messick (1988; 1989; 1996) haben zu einem Verständnis von Validität geführt, das bis heute in der Fachdiskussion um Sprachtests als Bezugspunkt gültig ist. Er hält die alles-oder-nichts-Sichtweise von Lado für falsch und betont hingegen, dass Tests einem Kontinuum zuzuordnen sind (*It is important to note that validity is a matter of degree, not all or none*; Messick, 1989: 13). Messick betont, dass die Validität nicht als Eigenschaft des Tests angesehen werden sollte. Er propagiert eine andere Sichtweise: Es gehe bei der Erhebung der Validität vielmehr darum, Argumente und Belege für die Zweckmäßigkeit und Rechtfertigung bestimmter Interpretationen der Testergebnisse zu finden.

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device per so but rather the inferences derived from test scores or other indicators (Cronbach, 1971) – inferences about score meaning or interpretation and about the implications for action that the interpretation entails (Messick, 1996: 245).

Dann ist die Bestimmung von Validität auch nicht zu einem bestimmten Zeitpunkt möglich, sie muss vielmehr über einen längeren Zeitraum als kontinuierlicher Prozess erfolgen. Ähnlich wie Messick argumentieren auch Bachman und Palmer:

> When we interpret scores from language tests as indicators of test takers' language ability, a crucial question is, 'To what extent can we justify these interpretations?' (Bachman/Palmer, 1996: 21). It is important for test developers and users to realize that test validation is an on-going process and that the interpretations we make of test scores can never be considered absolutely valid (1996: 22).

Außerdem nahm Messick eine Neubestimmung der Konstruktvalidität vor. Traditionell versteht man unter der Konstruktvalidität das Ausmaß, in dem ein Test einer Theorie über ein zugrunde liegendes Konstrukt entspricht. Sie ist nach dieser Vorstellung nur einer von mehreren Aspekten, welche die Validität eines Tests ausmachen. Andere sind beispielsweise: Inhaltsvalidität (*content validity*; Stellen die Aufgaben eine repräsentative Stichprobe aller Aufgaben aus dem zu prüfenden Sprachbereich dar?), Augenscheinvalidität (*face validity*; Halten Laien den Test mit Blick auf das Prüfungsziel für eine angemessene Messmethode?), Übereinstimmungsvalidität (*concurrent validity*; Korreliert der Test hoch mit Kriterien, deren Konstrukt bekannt ist?) oder Vorhersagevalidität (*predictive validity*; Wie gut sagt der Test künftige Leistungen im Bereich des Testkonstrukts voraus?). Messick argumentiert hingegen, dass diese verschiedenen Typen von Validität nicht nebeneinander stehen. Seiner Ansicht nach stellt die Konstruktvalidität ein übergeordnetes Merkmal dar, welches die übrigen einschließt:

> Because score meaning is a *construction* that makes theoretical sense out of both the performance regularities summarized by the score and its pattern of relationships with other variables, the psychometric literature views the fundamental issue as *construct* validity (Messick, 1996: 246. Hervorhebungen im Original).

Die Vorstellung Messicks von der Konstruktvalidität als dem beherrschenden Kriterium der Validität ist anerkannt und weit verbreitet. Unter einem Konstrukt versteht man eine angenommene Eigenschaft, wie z. B. eine sprachliche Fähigkeit, welche die Grundlage für einen Test darstellt und die als Indikator für Interpretationen der Ergebnisse gewertet wird (Davies *et al.*, 1999: 31). Messick (1996: 248-253) zählt sechs unterscheidbare Aspekte der Konstruktvalidität als übergeordnetes Kriterium für Validität auf: den Inhaltsaspekt (*content aspect*; Relevanz des Inhalts), den substanziellen Aspekt (*substantive aspect*; Übereinstimmung mit dem theoretischen Konstrukt), den strukturellen Aspekt (*structural aspect*; Angemessenheit der Messinstrumente mit der Struktur des Konstrukts), den Aspekt der Generalisierbarkeit (*generalizability aspect*; Repräsentativität des Messverfahrens mit Blick auf das Konstrukt), der externe Aspekt (*external aspect*; Übereinstimmung der Messergebnisse mit dem theoretischen Konstrukt) und der Aspekt der Auswirkungen (*consequential aspect*; Auswirkungen der Interpretationen).

Bachman und Palmer lehnen sich eng an das Modell von Messick an, folgen ihm jedoch nicht in allen Einzelheiten.

> The term *construct validity* is therefore used to refer to the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure. Construct validity also has to do with the domain of generalization to which our score interpretations generalize (Bachman/Palmer, 1996: 21).

Sie betonen ebenfalls die zentrale Rolle der Konstruktvalidität, gliedern jedoch einige Aspekte aus. Den Zusammenhang zwischen der Sprachverwendung im Test und der Sprachverwendung in "realen" Situationen, die mit dem Test erfasst werden soll, bezeichnen sie als Authentizität (*authenticity*) und stellen sie als eigenes Kriterium für die Nützlichkeit von Tests vor. Ebenso verfahren sie mit der Art und Weise, wie im Sprachtest die Sprachkompetenz der Testteilnehmer aktiviert wird, die sie als Interaktivität (*interactiveness*) bezeichnen. Ein weiterer Unterschied: die Ausgliederung der Testauswirkungen. Diese werden in der Testtheorie ebenfalls meistens der Validität zugerechnet. So verwendet Morrow den Ausdruck *washback validity*, Frederiksen und Collins sprechen von *systemic validity* und Messick spricht von *consequential validity* (Frederiksen/Collins, 1989; Messick, 1996; Morrow, 1986). An den besonderen Abgrenzungen wird deutlich, dass das Modell von Bachman und Palmer durchaus eigene Akzente setzt. Gemeinsam ist ihrem Ansatz mit anderen, dass Validität auf die Interpretation der Ergebnisse bezogen wird und nicht als Eigenschaft des Tests angesehen wird. Die Interpretationen sollen sich auf möglichst vielfältige Beobachtungen und Argumentationen stützen und nicht nur auf einen bestimmten Aspekt, der möglicherweise als im Vordergrund stehend angesehen wird (siehe Kapitel 2.1).

Bachman und Palmer sehen die so genannte Augenscheinvalidität (*face validity*) nicht als zentralen Bestandteil der Konstruktvalidität an. Unter der Augenscheinvalidität versteht man die Ansichten von Laien über die Qualität eines Sprachtests. Die Bedeutung der Augenscheinvalidität wird in der Literatur zu Sprachtests kontrovers diskutiert. Dabei lassen sich zwei Argumentationen ausmachen: Zum einen wird die Augenscheinvalidität als populärer, aber irrelevanter Aspekt der Validität angesehen. Eine tragfähige Basis für Schlussfolgerungen zur Validität wird in den Aussagen von Laien nicht gesehen. So argumentiert beispielsweise Stevenson (1983), der die Augenscheinvalidität als Populärvalidität (*pop validity*) karikiert. Die Vernachlässigung der Augenscheinvalidität im Modell von Bachman und Palmer lässt auf eine ähnliche Sichtweise schließen. Auf der anderen Seite sieht man in der Wahrnehmung der Testteilnehmer und der Test-

anwender durchaus eine wichtige Komponente der Validität. Wenn die Kandidaten einen Test für sinnvoll halten, strengen sie sich möglicherweise mehr an und rufen ihre volle Leistungsfähigkeit ab. Einige Forschungsergebnisse mit Blick auf die Testauswirkungen lassen die Bedeutung der Augenscheinvalidität als durchaus beachtenswert erscheinen. Watanabe (1996) beschreibt, wie die Vorbereitung auf (englischsprachige) Aufnahmeprüfungen zu japanischen Universitäten mit der Wahrnehmung des Sprachtests durch die Lehrkräfte und die Kandidaten zusammenhängt. Alderson und Hamp-Lyons (1996) beobachten den Unterricht von Lehrkräften, die sowohl TOEFL-Vorbereitungskurse als auch Sprachkurse ohne Prüfungsziel erteilen. Sie kommen zu dem Schluss, dass der tatsächliche Inhalt des Tests nicht die einzige Ursache für Auswirkungen auf die Lehr- und Lernprozesse darstellt. Auch die Ansichten der Lehrkräfte über den Test spielen eine Rolle. Die Beobachtungen von Lewkowicz (1996; zit. n. Alderson, 2000a: 29-30) verdeutlichen, dass die Kandidaten einen Test für den akademischen Sprachgebrauch ganz anders einschätzten als die Testentwickler. Sie schlussfolgert, dass die Authentizität des Sprachtests sehr stark von der Wahrnehmung der Kandidaten abhängt. Bei Sprachstandstests für den Hochschulzugang sind diese Überlegungen jedoch unwesentlich, da es Tests mit gewichtigen Konsequenzen sind. Kandidaten dürften unabhängig von ihrer Meinung über den Test ihre bestmögliche Leistung abrufen.

Aussagen über die Konstruktvalidität können gewonnen werden, indem die Testergebnisse mit den Ergebnissen anderer Tests korreliert werden, deren Konstrukt bekannt ist. Argumente zur Konstruktvalidität können auch durch eine inhaltliche Analyse erfolgen (Chapelle, 1999; Bachman, 1990).

Empirische Untersuchungen zur Validität des TestDaF und der DSH sind mir nicht bekannt. Es gilt jedoch: Ein ungenauer Test lässt keine sinnvollen Schlussfolgerungen gemäß der Messintention zu. Da die Reliabilität der DSH als gering einzuschätzen ist, muss auch von einer eingeschränkten Validität (mit Blick auf Sprachkenntnisse für das Studium) ausgegangen werden. Der standardisierte TestDaF erfüllt die Voraussetzung der Reliabilität in höherem Maße – Informationen zur Reliabilität werden jedenfalls systematisch erhoben (Eckes, 2003; Eckes/Grotjahn, in Druck).

Beide Tests sollen ein sehr komplexes Konstrukt erfassen, die Deutschkompetenz für ein Hochschulstudium. Eine hohe Validität ist eher zu beobachten, wenn sich das Test-

konstrukt gut erfassen lässt und präzise beschrieben wird. Es ist dann leichter, eine repräsentative Auswahl an Materialien und Aufgaben für den Test zu treffen. Das Konstrukt von Sprachtests für den Hochschulzugang ist jedoch keineswegs leicht zu erfassen. Die Sprachverwendungssituationen im Studium sind äußerst vielfältig, und sie unterscheiden sich auch von Studienfach zu Studienfach. Beim TestDaF verzichtet man ausdrücklich auf einen Bezug zu einem bestimmten Studienfach. Der TestDaF ist ein Test der allgemeinen Wissenschaftskommunikation (*Languages for general academic purposes*; siehe Kapitel 5.1, Seite 193 ff), bei dem die Beschreibung des Konstrukts und die Auswahl der Materialien ein besonderes Problem darstellt. Beim TestDaF könnte noch angeführt werden, dass der Test ähnlich vorgeht wie andere Sprachtests für den Hochschulzugang, bei denen man ebenfalls von einer hohen Konstruktvalidität ausgeht.

Die DSH wird dezentral von Personen erstellt, welche häufig Kontakt zu den Kandidaten haben. Die durchführenden Institute haben die Möglichkeit, eine Prüfung zu erstellen, welche das Studienfach berücksichtigt. Sie kann als Sprachtest mit Fachbezug konzipiert werden. Die adressatennahe Erstellung der Prüfung könnte ein Argument darstellen, welches die Existenz einer dezentralen Prüfung trotz ihrer Nachteile rechtfertigt. Dieser Gedanke wird in den Kapiteln 5 und 6 aufgegriffen.

Authenticity – Authentizität

Die folgenden drei Kriterien Authentizität, Interaktivität und Auswirkungen werden häufig zur Validität eines Sprachtests gezählt. Bachman und Palmer halten sie jedoch für so wichtig, dass sie eigens aufgeführt werden. Die Authentizität bezeichnet das Ausmaß, in dem der Test Sprachbenutzung außerhalb einer Testsituation widerspiegelt. Dies kann beispielsweise durch die Verwendung authentischer Texte geschehen, Texte also, welche nicht eigens für den Test verfasst wurden. Sprachtests mit einer hohen Authentizität tragen nach Bachman und Palmer zur Nützlichkeit von Sprachtests bei: Durch die Abbildung realer Sprachverwendungssituationen, die im Zusammenhang mit dem Testkonstrukt stehen, ermöglichen derartige Sprachtests eine verlässliche Interpretation der Ergebnisse. Hier wird eine ähnliche Argumentation wie bei direkten Performanztests verfolgt (siehe Kapitel 2.1, Seite 29). Außerdem gehen Bachman und Palmer davon aus, dass realitätsnahe Sprachtests von den Kandidaten positiv einge-

schätzt werden und sie durch den authentischen Sprachgebrauch möglicherweise zu besseren Leistungen angeregt werden.

Die Authentizität von Sprachtests ist kaum quantifizierbar oder empirisch nachweisbar. Ein Nachweis geschieht argumentativ, etwa durch eine inhaltliche Analyse der angestrebten Sprachverwendungssituation und der Testmaterialien. Man kommt dabei nicht zu einem alles-oder-nichts Ergebnis ("authentisch" oder "nicht authentisch"), sondern bestimmt einen Grad der Authentizität, der größer oder kleiner sein kann (Bachman/Palmer, 1996: 23-25).

Die Bedeutung des Kriteriums Authentizität ist umstritten. Lewkowicz (1997; zit. n. Grotjahn, 2000a: 319) weist darauf hin, dass ihrer Erfahrung nach auch zu Testzwecken erstellte Texte realitätsnah sein können, wodurch die tatsächliche Authentizität als Qualitätskriterium in den Hintergrund tritt. Außerdem spielt die Authentizität der Materialien für die Wahrnehmung der Testgüte durch die Kandidaten ihren Erfahrungen nach nur eine untergeordnete Rolle.

Bei der Konzeption des TestDaF lehnte man sich an die Vorgehensweise beim TOEFL an; der Authentizität wird keine große Rolle beigemessen. Im Falle des TOEFL nahm man eine Abwägung zwischen Konstruktvalidität und Authentizität vor. Chapelle, Grabe und Berns (1997: 26) argumentieren, dass angesichts der Vielfältigkeit der Sprachverwendungssituationen im Studium eine Authentizität nur scheinbar hergestellt werden kann. Wichtiger sei es, den Test so zu gestalten, dass gesicherte Interpretationen der Testergebnisse mit Bezug auf Sprachverwendungssituationen im Studium vorgenommen werden können, also Argumente für eine Konstruktvalidität. Dies bedarf ihrer Ansicht nach nicht unbedingt die Fokussierung auf eine Authentizität, die ohnehin kaum hergestellt werden kann. Ein weiteres Argument wird von Grotjahn vorgebracht: Da Sprachtests für den Hochschulzugang schwer wiegende Konsequenzen haben, bei denen die Aspekte der Interaktivität und der Authentizität für die Kandidaten eine untergeordnete Rolle spielen, sei zu erwarten, "dass die Probanden versuchen, auch Aufgaben, die von ihnen als wenig authentisch wahrgenommen werden, möglichst optimal zu lösen" (Grotjahn, 2000a: 319).

Die Rahmenordnung der DSH ermöglicht es, einen besonderen Testzuschnitt zu entwerfen, das heißt, sie kann an den speziellen Bedürfnissen von Testanwendern oder

Testteilnehmern ausgerichtet werden. Es können Materialien und Texte verwendet werden, die einen hohen Grad an Authentizität herstellen. Durch die Möglichkeit, in einzelnen Prüfungsteilen der DSH auch einen Fachbezug herzustellen, ist eine besonders realitätsnahe Gestaltung der DSH möglich (siehe Kapitel 5 und 6).

Interactiveness – Interaktivität

Bachman und Palmer verstehen unter *interactiveness* das Ausmaß, in dem die Sprachfähigkeit der Kandidaten, ihr Wissen über das Thema und ihr Interesse an der Erfüllung der Aufgabe einbezogen werden. Da der deutsche Begriff "Interaktivität" eher an Computerspiele denken lässt, wäre eine beschreibende Übertragung etwa mit "Grad der Aktivierung und Einbeziehung der Kandidaten" passender. Wegen seiner Prägnanz übernehme ich jedoch den Begriff "Interaktivität".

Das Konzept der Interaktivität beruht auf der Vorstellung von Bachman und Palmer vom Wesen der Sprachkompetenz (siehe Kapitel 2.1, Seite 35). Wenn man beispielsweise der strategischen Kompetenz als Vermittlerin zwischen den einzelnen Komponenten eine zentrale Rolle für die Sprachkompetenz zuweist, ist es nur folgerichtig, diese Kompetenz auch in Sprachtests zu aktivieren. Ähnlich wie die Authentizität gilt auch für die Interaktivität, dass sie kein konstitutives Merkmal der Nützlichkeit darstellt. Mit Blick auf einen bestimmten Einsatz können Sprachtests nützlich sein, ohne besonders authentisch oder besonders interaktiv zu sein. Andererseits sollten diese Merkmale nicht aus den Augen verloren werden, da sie eng mit dem Modell des Konstrukts der Sprachkompetenz zusammenhängen. Die Forderung nach einer hohen Interaktivität steht in Einklang mit theoretischen Prämissen über den gesteuerten Spracherwerb, welche eine Teilnehmerorientierung in Sprachkursen und die Einbindung der Kursteilnehmer in aktive Lernprozesse fordern. Mit dem Konzept weisen Bachman und Palmer auf ein Kriterium für die Nützlichkeit von Sprachtests hin, das durch die Betonung von kommunikativen Lernprozessen an Bedeutung gewonnen hat.

Der Grad der Aktivierung und Einbeziehung der Testteilnehmer lässt sich durch eine inhaltliche Analyse, durch Beobachtung der Teilnehmer und durch eine Analyse der Testergebnisse nachweisen. Ebenso wie bei dem Kriterium der Authentizität kommt man zu einem beschreibenden Ergebnis.

Beim Vergleich der DSH und des TestDaF mit Blick auf das Kriterium der Interaktivität fällt der Prüfungsteil zum Sprechen auf. Durch die Verwendung des Kassetten gesteuerten Testformats beim TestDaF entsteht eine besondere Kommunikationssituation (Kenyon, 2000). Die DSH sieht ein Prüfungsgespräch vor. Ein weiterer Unterschied, der Auswirkungen auf den Grad der Interaktivität haben könnte, liegt in dem Fachbezug. Wenn in der DSH ein Fachbezug zum zukünftigen Studienfach hergestellt wird, könnte das zu einer intensiveren Auseinandersetzung mit der Aufgabenstellung führen, als wenn ein beliebiges Thema gewählt wird. Abgesehen davon erkenne ich keine unterschiedlichen Merkmale zwischen den beiden Prüfungen, welche Auswirkungen auf die Interaktivität erwarten lassen.

Impact – Auswirkungen

Auswirkungen von Prüfungen werden häufig als ein Naturgesetz angesehen. Eine Rückwirkung von einer Prüfung auf die Vorbereitung darauf erscheint in der Tat plausibel. Andererseits verständigt man sich selten über den genauen Bedeutungsumfang des Begriffs: Welche Auswirkungen sind gemeint? Auswirkungen auf die Lernenden, auf die Lehrenden, auf die Einstellung zum Lernen, auf die Inhalte, auf die Methoden, auf das Curriculum, auf die Gestaltung der Lehrwerke? Sind die Auswirkungen auf alle Lerner, auf alle Lehrer gleich stark? Bachman und Palmer präzisieren den Begriff, indem sie bei den Auswirkungen einer Prüfung zwischen der Mikro- und der Makroebene unterscheiden. Die Mikroebene bezieht sich auf die Lehr- Lernprozesse einzelner Personen, bei der Makroebene geht es um die Auswirkungen auf die Gesellschaft und das Bildungssystem.

In der englischsprachigen Forschung spricht man neben *impact* auch von *washback* oder *backwash*. *Washback* und *backwash* werden unterschiedlich abgegrenzt: Einige Autoren beziehen *backwash* auf die allgemeine Pädagogik und *washback* auf die Sprachen, bisweilen wird *washback* auf die britische Angewandte Linguistik eingeschränkt. Manchmal werden beide Begriffe auch synonym verwendet. Bachman und Palmer (1996: 30-31) verwenden *washback* und schränken dies auf die Mikroebene ein. Dies entspricht der Definition von Hughes (2003: 1), der *washback* bestimmt als: "*the effect of testing on teaching and learning*". Die Definition Cohens (1994: 41) ist weiter. Er bezieht die

Auswirkungen von Prüfungen allgemein auf pädagogische Maßnahmen und Ansichten ("*how assessment instruments affect educational practices and beliefs*"). In der deutschsprachigen Forschung greift man häufig auf die englischen Begriffe zurück oder spricht vom "Backwash-Effekt". Auch deutsche Termini wie etwa "Rückkopplungsphänomen", "Rückwirkungsmechanismus" oder "Rückstromeffekt" werden verwendet (Alderson/Wall, 1993; Bachman/Palmer, 1996; Bailey, 1996; Cheng/Watanabe/Curtis, 2004; Grotjahn, 2000a; Hughes, 2003; Schifko, 2001; Wall, 2000). Ich lehne mich an den Sprachgebrauch von Bachman und Palmer an und spreche von "Auswirkungen".

Bei Sprachprüfungen für den Hochschulzugang geht man von deutlichen Auswirkungen auf die Lehr- und Lernprozesse aus. "Diese Prüfung [die DSH] hat großen Einfluss auf den Lehr-Lern-Prozess in der Prüfungsvorbereitung" (Lee, 1998; 4). Stützen lässt sich die Einschätzung durch die hohe Bedeutung der Sprachtests für den Hochschulzugang.

> If a test is regarded as important, if the stakes are high, preparation for it can come to dominate all teaching and learning activities (Hughes, 2003: 1);
>
> Je höher der Status und je gewichtiger die Konsequenzen eines Tests, desto stärker ist sein Backwash-Effekt (statusbezogener Aspekt) (Schifko, 2001: 830).

Sowohl der TestDaF als auch die DSH sind Tests mit gewichtigen Konsequenzen. Beide haben Auswirkungen auf die Testvorbereitung. Durch die Unterschiede könnten sich auch unterschiedliche Auswirkungen ergeben:

- Interessant sind die Auswirkungen der unterschiedlichen Formate: Dazu gehören die unterschiedliche Vorgehensweise bei der Mündlichen Prüfung oder die unterschiedliche Anzahl der Texte beim Leseverstehen. Hier steht eine Bewertung der Auswirkungen noch aus. Der DSH-Grammatiktest gehört zu den Besonderheiten der DSH, der auch nach der Überarbeitung der Rahmenordnung in verkleinertem Umfang enthalten ist. Führt der Grammatiktest zu einer einseitigen Konzentration auf die strukturelle Seite der Sprache? Studien zu dieser Frage werden in Kapitel 4.4 (Seite 158 ff) vorgestellt.

- Ein weiterer Unterschied: Bislang konnte die DSH nur einmal wiederholt werden, der TestDaF jedoch beliebig oft. Doch hier sieht die neue Rahmenordnung vor, dass auch die DSH unbegrenzt oft wiederholt werden kann (HRK/KMK, 2004). Dies dürfte in der Praxis immer schon möglich gewesen sein, aber das Damoklesschwert

der nur einmaligen Wiederholung, die offiziell den Kandidaten mitgeteilt wurde, wurde von vielen als besonderer Druck empfunden.

- Beim TestDaF dürften die Kosten zu einer hohen Augenscheinvalidität beitragen. Auch für die DSH wird von vielen Ausrichtern eine Gebühr erhoben, sie liegt in der Regel aber deutlich unter der Gebühr für den TestDaF.

- Die DSH ist als dezentral ausgerichtete Prüfung häufig eng an einen Vorbereitungskurs gebunden. Daher dürfte die DSH die Auswirkung haben, dass Kandidaten an demjenigen Institut einen Sprachkurs belegen, das auch für die DSH verantwortlich ist. Die institutionelle Anbindung der DSH an Hochschulen führt dazu, dass auch die Vorbereitung häufig in einem universitären Umfeld stattfindet, was sicherlich positiv ist, auch wenn eine Anbindung an den Studienalltag mit den damit verbundenen informellen Lerngelegenheiten nicht immer gelingen dürfte. Daher bemüht sich das TestDaF-Institut folgerichtig, möglichst viele Hochschulen als Prüfungszentren zu gewinnen.

- Ein gewichtiger Unterschied zwischen beiden Prüfungen ist, dass der TestDaF auch im Heimatland absolviert werden kann. Die Anmeldung zu einer DSH ermöglicht die Einreise nach Deutschland und die Studienvorbereitung im Zielland. Wenn die sprachliche Vorbereitung erfolgreich ist und die Studienbewerber einen Studienplatz erhalten, ist die Vorbereitung in Deutschland sicherlich ein Vorteil. Ist das Studium in Deutschland jedoch nur eine Option unter mehreren, ist die Möglichkeit, den Test im Heimatland abzulegen, sicherlich günstiger.

Insgesamt scheint eine große Zurückhaltung bei Aussagen zu Auswirkungen dieser Tests angebracht. Lassen sich angesichts der Komplexität des Phänomens überhaupt sinnvolle Aussagen zu den Auswirkungen von Sprachprüfungen treffen? Zweifel sind angebracht, denn im Fall der Sprachprüfungen für den Hochschulzugang kommt hinzu, dass die Zielgruppe sehr heterogen ist; die Lernwege der Prüfungsteilnehmer sind sehr unterschiedlich.

Practicality – Ökonomie

Mit dem Kriterium der Testökonomie verweisen Bachman und Palmer darauf, dass ein zweckmäßiger Test auch in einem sinnvollen Verhältnis zum Aufwand stehen muss. Sie definieren "*practicality*" als das Verhältnis zwischen den Mitteln, die für die Konzeption, Entwicklung und den Einsatz des Tests benötigt werden und den Mitteln, die dafür zur Verfügung stehen. Zu den Mitteln, die bei Sprachtests zu berücksichtigen sind, zählen sie Personal, Material (Räume, Ausrüstungen, Unterlagen) und Zeit (für die Entwicklung/Auswertung und für die Durchführung). Auch hier möchte ich die Wahl der Übersetzung erläutern: Während das Adjektiv "praktikabel" den Begriff möglicherweise noch angemessen erfasst, halte ich das Substantiv "Praktikabilität" für unpassend. Gängiger und passender ist wohl der Begriff der "Ökonomie", denn schließlich geht es Bachman und Palmer ausdrücklich um ein Verhältnis zwischen ökonomischen Größen.

In der Theorie lässt sich die Ökonomie eines Tests rechnerisch bestimmen: Man addiere die zur Verfügung stehenden Ressourcen und teile sie durch die benötigten Ressourcen. Ist der Quotient größer oder gleich eins, so ist der Test ökonomisch. Diese Rechnung trifft in der Praxis auf Schwierigkeiten, weil selten ein festes Budget oder Zeitkonto zur Verfügung steht. Das Anliegen ist vielmehr, den Test so ökonomisch wie möglich zu machen. Die verwendeten Ressourcen sind dabei nicht immer quantifizierbar, sondern beruhen auf subjektiven Einschätzungen. Wie viele Seiten die Testunterlagen für die Testteilnehmer maximal umfassen sollten, hängt schließlich nicht allein vom Budget ab, sondern von der Beurteilung der jeweiligen Situation.

Bei dem Vergleich der Ökonomie zwischen dem TestDaF und der DSH stellt sich eine grundsätzliche Frage: Sind zentrale oder dezentrale Sprachtests ökonomischer? Die Testentwicklung ist bei zentralen Tests sicherlich ökonomischer, die Durchführung und Korrektur kann vor Ort einfacher und häufig schneller geschehen. Der größte Vorteil dezentraler Tests liegt für Testkandidaten und Testanwender in der hohen Flexibilität. Es ist möglich, die DSH für eine kleine Gruppe, ja sogar für einzelne Kandidaten an einem kurzfristig festgesetzten Termin abzuhalten. Dies ist nicht ökonomisch, aber außerordentlich praktisch. In der hohen Flexibilität dürfte der Hauptgrund dafür liegen, dass die DSH trotz mangelhafter Reliabilität weiterhin geschätzt wird.

Auch zum Bereich der Testökonomie im weiteren Sinne gehört, dass das Testergebnis klar und verständlich ausgewiesen werden soll und dass die Vorgehensweise der

Prüfung vermittelbar sein soll. In dieser Hinsicht gibt es bei beiden Tests Schwächen: Der Ergebnisausweis beim TestDaF ist aus der Sicht der Testtheorie zwar unangreifbar, der Umgang damit für Kandidaten und Testanwender jedoch mühsam: Warum werden beispielsweise beim TestDaF Niveaustufen ausgewiesen, die sich nur mit Mühe etablierten Skalen wie denen des Europarats (vgl. Europarat/Rat für kulturelle Zusammenarbeit, 2001) zuordnen lassen? Die Anwender sind gezwungen, sich mit den neuen TestDaF-Ergebnisklassen auseinanderzusetzen und den Umgang damit zu lernen. Bei der DSH ist seit der Überarbeitung der Rahmenordnung (HRK/KMK, 2004) eine äußerst verwirrende Ermittlung des Gesamtergebnisses vorgesehen. Zum einen sollen Schriftliche und Mündliche Prüfung im Verhältnis 70:30 gewichtet werden, zum anderen kann eine bestimmte Ergebnisklasse nur ausgewiesen werden, wenn sie sowohl in der Schriftlichen als auch in der Mündlichen Prüfung erzielt wird. In Einzelfällen ist die Folge: Obwohl insgesamt eine Prozentzahl ermittelt wird, die *über* dem Schwellenwert für eine Ergebnisklasse liegt (z. B. 67 % für DSH-2), muss in Einzelfällen als Prüfungsergebnis die *niedrigere* Ergebnisklasse (DSH-1) ausgewiesen werden, wenn dieser Wert (67 %) nicht in Mündlicher und Schriftlicher Prüfung erzielt wurde. Außerdem verwirrend: Die Anzahl der Prüfungsteile entspricht nicht der Anzahl der Ergebnisse, die auf dem Zeugnis ausgewiesen werden sollen. Die Schriftliche Prüfung besteht aus drei Prüfungsteilen, ausgewiesen werden aber vier Ergebnisse, da ein Prüfungsteil aus zwei Teilprüfungen besteht, die außerdem unterschiedlich gewichtet werden. Nicht gelöst wurde die Behandlung der Mündlichen Prüfung: Wenn keine Mündliche Prüfung stattfindet, soll ein fiktiver Wert in die Berechnung einbezogen werden. Dies sind überflüssige Stolpersteine, welche die Ermittlung des Ergebnisses verkomplizieren und zu einem hohen Erklärungsbedarf führen. Wir machen es – so lautet die Botschaft von derartigen Regelungen – euch nicht leicht! Erst müssen ausländische Studienbewerber die deutsche Sprache meistern, dann sollen sie sich auch mit komplizierten Testergebnissen auseinandersetzen.

Durch den DSH-Grammatiktest, bei dem die sprachliche Richtigkeit das einzige Bewertungskriterium darstellt, verfügt die DSH über einen Prüfungsteil mehr als der TestDaF, was Fragen nach seiner Berechtigung weckt. Einige Sprachtests für den Hochschulzugang kommen ohne einen Prüfungsteil zur Grammatik aus, ein Verzicht würde auch ein Gewinn an Ökonomie bedeuten (siehe Kapitel 3.2, Seite 93).

Zusammenfassung

Eine Analyse der DSH nach Nützlichkeitskriterien von Sprachtests kommt zu einem negativen Ergebnis, weil das zentrale Kriterium, die Reliabilität, nicht erfüllt wird. Wenn es zu einer hohen Reliabilität an einem Institut kommt, ist das der Leistung einzelner Prüfungsbeauftragter zuzuschreiben. In der Regel ist bei der nichtstandardisierten DSH von einer geringen Reliabilität auszugehen, was von den Kandidaten auch wahrgenommen wird: Sie melden sich an verschiedenen Orten für die DSH an. Eine weitere Folge: Interpretationen mit Blick auf das Testkriterium sind fragwürdig. Der TestDaF dürfte demgegenüber das Kriterium der Reliabilität zu einem hohen Grad erfüllen. Damit ist eine wichtige Voraussetzung für die Konstruktvalidität erfüllt.

Andererseits eröffnet die dezentrale Ausrichtung der DSH Spielräume, welche genutzt werden können. Eine DSH kann einen höheren Grad an Authentizität oder Interaktivität aufweisen. Beim TestDaF konzentriert man sich auf die Sicherstellung einer hohen Zuverlässigkeit, Kriterien wie Authentizität oder Interaktivität wird keine große Bedeutung beigemessen.

Welche besonderen Auswirkungen auf die Prüfungsvorbereitung sind von der DSH zu erwarten? Zu nennen sind meiner Ansicht nach der Grammatiktest und die institutionellen Besonderheiten einer jeden DSH, welche eine Vorbereitung auf eine Prüfung an einem bestimmten Institut nahe legt. Neben der hohen Flexibilität ist die hochschulnahe Vorbereitung der größte Vorteil der DSH, denn sie führt auch zu vielen informellen Lerngelegenheiten zur Vorbereitung auf ein Fachstudium. Beim TestDaF ist demgegenüber zu fragen, ob der Verzicht auf den Grammatiktest zu einer Vernachlässigung der sprachlichen Richtigkeit in studienvorbereitenden Sprachkursen führt. Die Möglichkeit, den TestDaF in Testzentren im Heimatland absolvieren zu können, hat sicherlich Vorteile für die Planung des Auslandsstudiums. Mit dem Sprachtest in der Tasche verfügen ausländische Studienbewerber bereits vor der Einreise nach Deutschland über eine höhere Planungssicherheit. Ob den ausländischen Studienbewerbern die informelle Eingewöhnung an deutsche Hochschulen fehlt, muss sich noch zeigen.

3. Grammatik in Sprachtests

Übersicht: Kapitel 3
Im Kapitel "Grammatik in Sprachtests" geht es um den Umgang mit der strukturellen Seite der Sprache in einigen Sprachtests für den Hochschulzugang. Am Beginn des Kapitels stehen allgemeine Überlegungen zur Behandlung von Grammatik in Sprachtests. In Kapitel 3.1 wird der Umgang mit Grammatik in ausgewählten deutschen und englischen Sprachtests vorgestellt, in Kapitel 3.2 stelle ich Fragen an Grammatiktests aus sprachwissenschaftlicher und testmethodischer Sicht.

Nachdenken über Grammatiktests ist nicht in Mode. Waren Grammatiktests in den 1960er Jahren noch ein selbstverständlicher Bestandteil zahlreicher Sprachtests, scheint die Attraktivität von Grammatiktests inzwischen deutlich nachgelassen haben. Von Interesse ist das Nachdenken über Grammatiktests im Zusammenhang mit deutschen Sprachtests für den Hochschulzugang, weil die formale Seite der Sprache bzw. sprachliche Korrektheit unterschiedlich behandelt wird. Während man beim TestDaF und anderen Sprachtests für den Hochschulzugang auf einen Grammatiktest verzichtet, enthält die DSH einige Items zur Grammatik. Bis zur Überarbeitung der DSH 2004 war sogar ein eigenständiger Prüfungsteil zur Grammatik enthalten. In der überarbeiteten DSH-Rahmenordnung wird jedoch gefordert, dass der Umfang des Grammatiktests verringert werden soll, was ich als Ausdruck eines gewissen Unbehagens deute.

Ausgehend von einem psychometrisch-strukturalistischen Sprachverständnis stand in den 1960er Jahren das Testen isolierter sprachlicher Elemente, zu denen neben Phonetik und Lexik auch die Grammatik zählte, nicht in Frage (Lado, 1961; 1971). Die Sprach-

kompetenz ergab sich aus der Summe der Komponenten. Das zentrale Anliegen war die Gewährleistung einer hohen Reliabilität, welche u. a. durch die Verwendung von erprobten Multiple-Choice Items erzielt wurde. Einzelne Items oder ganze Prüfungsteile zur Grammatik gehörten einfach dazu. Seit dem Ende der 1970er bzw. Anfang der 1980er Jahre richtet sich die Aufmerksamkeit verstärkt auf das integrierte Testen der sprachlichen Fertigkeiten Lesen, Hören, Schreiben und Sprechen. Der kommunikative Ansatz stellte auch bei Sprachtests an Materialien und Aufgaben den Anspruch, authentisch zu sein. Wenn Grammatik geprüft wird, dann als Bestandteil komplexer, sprachlicher Äußerungen in Performanztests unter der Fragestellung: Hat der Kandidat die (authentische) Sprachverwendungssituation angemessen gelöst? Man erkannte an, dass eine isolierte Bewertung der sprachlichen Elemente, welche den sprachlichen Äußerungen möglicherweise zugrunde liegen, weder exakt möglich noch wirklich erforderlich ist (zur Diskussion um kommunikative Sprachtests siehe Canale, 1984; Fulcher, 2000; Morrow, 1979; Rea-Dickins, 1991; Weir, 1990; Widdowson, 2001).

Eine Übersicht über die Prüfungsteile in einigen Sprachprüfungen mit und ohne Grammatiktest geht aus Tabelle 5 hervor (Seite 66). In neu entwickelten Tests wird dem Testen der integrierten Fertigkeiten häufig der Vorzug gegeben. In der Entscheidung über das Testformat spiegelt sich ein bestimmtes Sprachverständnis; sie dürfte sich nicht allein an der wissenschaftlichen Diskussion orientieren, sondern auch von Marktinteressen und Gewohnheiten geleitet sein. Eine Ausnahme bildet IELTS, dessen Format als Folge einer umfangreichen Studie verändert wurde: Ein Grammatiktest, der Bestandteil einer früheren Version des IELTS war, wurde gestrichen (Alderson, 1993). In der aktuellen Version enthält IELTS – wie auch der TestDaF – keinen Grammatiktest, sondern besteht aus insgesamt vier Prüfungsteilen: Hörverstehen, Leseverstehen, Schriftlichem Ausdruck und Mündlicher Prüfung. Wie ist es dazu gekommen? In einer IELTS-Pilotversion war ein Grammatiktest enthalten. Es wurde Wert darauf gelegt, dass sich die Anforderungen der einzelnen Testteile nicht überschneiden. Der Grammatiktest bestand aus Aufgaben nicht nur zu grammatischen Strukturen, sondern auch zur Textkohärenz und zum Wortschatz. Die Kontextualisierung legte dabei die Auswahl der getesteten Phänomene fest. Eine Untersuchung der Teilergebnisse aus umfangreichen Probedurchläufen führte zu der Erkenntnis, dass die Korrelation zwischen den Ergebnissen aus dem Grammatiktest und den Testteilen Leseverständnis und Hörverständnis hoch war. Ob die Korrelation auch bei weniger kontextualisierten Auf-

gabentypen in gleichem Maße bestanden hätte und ob ein anders gearteter Grammatiktest zusätzliche Informationen geboten hätte, wurde nicht überprüft. Allerdings hat man die Korrelation mit verschiedenen Leseverständnistests feststellen können. In Bezug auf die Reliabilität stellte man fest, dass ein Verzicht auf den Grammatiktest nicht zu Abstrichen bei der Nützlichkeit führte. Dem Grammatiktest wurde zwar eine hohe Reliabilität und auch eine hohe Validität in Bezug auf das Testziel "Sprachkompetenz für die Aufnahme eines Studiums" bescheinigt, doch da er die Ergebnisse anderer Prüfungsteile offensichtlich nur bestätigte, wurde er gestrichen. Diese Maßnahme zur Verbesserung der Testökonomie hat bis heute Bestand (Alderson, 1993; Clapham, 2000).

War IELTS zunächst der kleine Bruder des TOEFL, so scheint sich das Verhältnis umgekehrt zu haben. Jedenfalls findet derzeit eine grundlegende Überarbeitung des TOEFL statt. Im Ergebnis wird der TOEFL dem IELTS (und dem TestDaF) ähneln. Auch der TOEFL wird vier Prüfungsteile zu den sprachlichen Fertigkeiten enthalten, und man wird auf einen Grammatiktest verzichten.

Ganz im Trend liegt demnach der TestDaF. Die Verwendung eines Prüfungsteils zur Grammatik wurde offensichtlich zu keiner Zeit erwogen, begründet wurde der Verzicht jedenfalls nicht. Die Auswahl der Prüfungsteile hat möglicherweise Auswirkungen: Mit der Entscheidung, auf einen Grammatiktest zu verzichten, setzt man ein Signal für die Bedeutung sprachlicher Richtigkeit und ein Signal für den Stellenwert von Grammatik im Sprachunterricht.

3.1. Grammatik in Sprachtests für den Hochschulzugang: Beispiele

Übersicht: Kapitel 3.1

In diesem Kapitel stelle ich vor, wie Grammatik in Sprachtests für den Hochschulzugang behandelt wird. Dazu wähle ich neben der DSH und dem TestDaF folgende Tests aus: "Test of English as a Foreign Language" (TOEFL), "Cambridge Certificate of Proficiency Examination" (CPE), "Zentrale Oberstufenprüfung des Goethe-Institus" (ZOP), "Kleines Deutsches Sprachdiplom" (KDS) und " International English Language Testing System" (IELTS).

Tabelle 5 bietet eine Übersicht über einige Sprachtests für den Hochschulzugang. Vor allem Tests, die erst vor einigen Jahren konzipiert wurden, enthalten keinen expliziten Grammatiktest.

Grammatik in Sprachtests ohne expliziten Grammatiktest

Werden Grammatikkenntnisse im TestDaF und im IELTS-Test nicht berücksichtigt? Welche Rolle spielt die grammatische Korrektheit bei diesen Tests? Die sprachliche Richtigkeit wird zwar nicht in einem eigenen Prüfungsteil, wohl aber im produktiven Prüfungsteil "Schriftlicher Ausdruck" (TestDaF) bzw. "Academic Writing Module" (IELTS) berücksichtigt. Die Kandidaten erhalten zum IELTS-Test folgende Informationen:

> Your writing will be assessed on your ability to:
> - organise, present and compare data
> - describe the stages of a process
> - describe an object or event
> - explain how something works
>
> You will also be judged on your ability to:

- answer the question without straying from the topic
- write in a way which allows your reader to follow your ideas
- use English grammar and syntax accurately
- use appropriate language in terms of register, style and content
  (Jakeman/McDowell, 1996: 4).

Ganz in der Tradition der kommunikativen Sprachtests steht die Bewältigung der authentischen Sprachverwendungssituation bei der Bewertung an erster Stelle. Doch der angemessene Gebrauch der englischen Grammatik wird nicht nur als Voraussetzung für die Sprachproduktion angesehen, sondern ausdrücklich als Bewertungskriterium genannt. Ähnlich verhält es sich beim TestDaF. Die Prüferinnen und Prüfer sollen sich bei der Bewertung des "Schriftlichen Ausdrucks" an folgenden Kriterien orientieren (Beispiel für die TDN 5, der höchsten Ergebnisklasse):

Tabelle 5: Beispiele für Sprachtests mit und ohne Grammatiktest

Sprachtests ohne Prüfungsteil zur Grammatik
Test Deutsch als Fremdsprache für ausländische Studienbewerber (TestDaF) 1. Mündlicher Ausdruck; 2. Leseverstehen; 3. Hörverstehen; 4. Schriftlicher Ausdruck
Österreichisches Sprachdiplom Deutsch: Mittelstufe 1. Lesen; 2. Hören; 3. Schreiben; 4. Sprechen
International English Language Testing System (IELTS) 1. Listening; 2. Reading; 3. Writing; 4. Speaking
Sprachtests mit Prüfungsteil zur Grammatik (Grammatiktests kursiv)
Deutsche Sprachprüfung für den Hochschulzugang (DSH) 1. Hörverstehen; 2. Leseverstehen und *wissenschaftssprachliche Strukturen*; 3. Textproduktion; 4. Mündliche Prüfung
Test of English as a Foreign Language (TOEFL) 1. Listening Comprehension; 2. *Structure and Written Expression;* 3. Reading Comprehension; zusätzliche Wahlelemente: Test of Written English (TWE), Test of Spoken English (TSE)
Certificate of Proficiency in English (CPE) 1. Reading; 2. Writing; 3. *Use of English;* 4. Listening; 5. Speaking
Zentrale Oberstufenprüfung (ZOP) 1. Mündliche Prüfung; 2. Texterklärung eines Lesetextes; 3. Hörverstehen; 4. Aufsatz; 5. *Ausdrucksfähigkeit*
Kleines Deutsches Sprachdiplom (KDS) 1. Mündliche Prüfung; 2. Erklärung eines Textes nach Inhalt und Wortschatz; 3. Diktat; 4. Aufgaben zur Lektüre; 5. *Aufgaben zur Überprüfung der Ausdrucksfähigkeit*

Bewertungskriterien für den Teiltest Schriftlicher Ausdruck, TDN 5:
Gesamteindruck
1. Der Text liest sich durchgängig flüssig.
2. Der Gedankengang kann problemlos nachvollzogen werden.
3. Der Text ist klar strukturiert.

Behandlung der Aufgabe
Der Text wird der Aufgabenstellung inhaltlich gerecht:
4. Alle in der Aufgabenstellung genannten Punkte werden in ausreichendem Umfang behandelt.
5. Die Informationen der Grafik(en) werden zusammengefasst; sie werden klar und folgerichtig dargestellt.
6. Im argumentativen Teil wird sachlich und ausführlich genug begründet und ggf. werden Beispiele als Belege angeführt.

Sprachliche Realisierung
Die sprachliche Realisierung ist der Aufgabenstellung angemessen:
7. Der Text hat
   - ein breites Spektrum an kohäsionsstiftenden Mitteln
   - ein breites Spektrum an syntaktischen Strukturen.
8. Der Wortschatz ist weitgehend differenziert und präzise.
9. Der Text enthält vereinzelt morphosyntaktische, lexikalische und orthografische Fehler (TestDaF-Institut, 2002).

Die Aspekte "Gesamteindruck", "Behandlung der Aufgabe" und "Sprachliche Realisierung" haben gleiches Gewicht. Bei der "Sprachlichen Realisierung" werden Textbezüge, Satzbau und Morphologie erwähnt. Auch beim "Gesamteindruck" dürfte die strukturelle Korrektheit bei der Bewertung berücksichtigt werden. Insgesamt fließen Grammatikkenntnisse mit maximal einem Drittel in die Bewertung dieses Prüfungsteils ein. Auch bei der Textproduktionen der DSH ist die grammatische Korrektheit ein Kriterium für die Bewertung. Das "DSH-Handbuch für Prüferinnen und Prüfer" schlägt bei der "Vorgabenorientierten Textproduktion" eine Berücksichtigung der Sprachrichtigkeit mit einem Anteil in Höhe von 30 Prozent bei der Bewertung vor (FaDaF, 2001: 6/9).

Abgesehen von einer *bewussten* Berücksichtigung der grammatischen Korrektheit und der strukturellen Vielfalt, wie sie in den produktiven Prüfungsteilen vorgesehen ist, fließt sie auch *unbewusst* in die Bewertung produktiver Prüfungsteile ein. McNamara stellte in einer Auswertung der Vorgehensweise von Prüfern fest, dass die grammatische Korrektheit bei der Bewertung von produktiven sprachlichen Äußerungen grundsätzlich eine große Rolle spielt. Er fasst zusammen:

> It seems, then, that a rater's perception of the grammatical and lexical accuracy of a candidate's performance is the most significant factor in the allocation of the candidate's total score (McNamara, 1996: 218).

Diese Beobachtung dürfte nicht nur auf die produktiven Prüfungsteile des TestDaF und des IELTS-Tests, sondern auch auf die produktiven Teile der DSH und anderer Sprachprüfungen für den Hochschulzugang zutreffen.

Zusammenfassend ist festzuhalten, dass grammatische Korrektheit nicht nur in Tests mit explizitem Grammatiktest in die Bewertung einbezogen wird. Wenn – wie in der DSH – zusätzlich ein Testteil existiert, in dem die sprachliche Korrektheit das einzige Bewertungskriterium darstellt, erhält die Grammatik freilich ein stärkeres Gewicht in der Gesamtbewertung.

Grammatik im "Test of English as a Foreign Language" (TOEFL)

Der wohl bekannteste Sprachtest der psychometrisch-strukturalistischen Phase, der "*Test of English as a Foreign Language*" (TOEFL), enthält derzeit noch einen Testteil "*Structure and Written Expression*", welcher aus zwei Sektionen besteht: "*Sentence Completion*" und "*Error Identification*" (siehe Abbildung 3). Beide Sektionen verwenden ein Multiple-Choice Format, auf eine Kontextualisierung der Items wird verzichtet. Der TOEFL ist der einzige der vorgestellten Tests, in dem auch Fehlererkennung zum Testgegenstand gemacht wird. Mit nur 15 bis 20 Minuten Bearbeitungszeit nimmt der Prüfungsteil einen vergleichsweise kleinen Raum ein. Die Grammatiktests im CPE, im KDS und in der ZOP dauern länger (90 Minuten) und sind in Bezug auf das Aufgabenspektrum umfangreicher.

Sentence Completion: Click on the one word or phrase that best completes the sentence.

The columbine flower, _____ to nearly all of the United States, can be raised from seed in almost any garden.

- native
- how native is
- how native is it
- is native

Error Identification: Click on the one underlined word or phrase that must be changed for the sentence to be correct.

One of the <u>most</u> difficult problems <u>in understanding</u> sleep is <u>determining</u> what the functions of sleep <u>is</u>.

(Quelle: ETS, Ohne Jahr)

Abbildung 3: TOEFL – Structure and Written Expression (Auszug)

Grammatik im "Cambridge Certificate of Proficiency Examination" (CPE)

Der Testteil "*Use of English*" im CPE besteht aus vier Sektionen. In Sektion 1 und 3 müssen Lücken ausgefüllt werden, nur in der ersten Sektion in einem zusammenhängenden Text, in der dritten Sektion in inhaltlich unverbundenen Sätzen. Die Sektionen 2 und 4 enthalten Transformationsaufgaben. Vorgegebene Wörter müssen umgeformt werden, damit sie in einen Text passen (Sektion 2). In der Sektion 4 sollen vorgegebene Sätze unter Gebrauch eines bestimmten Wortes verändert werden (siehe Abbildung 4).

Part 1: Read the text below and think of the word which best fits each space. Use only one word in each space.

The age of a garden *has* a great effect on the abundance of its wildlife. Since most animals depend ultimately on plants for their food, animal life cannot easily establish ..... in the absence of plant life. [...]

Part 2: Read the text below. Use the word given in capitals at the end of some of the lines to form a word that fits in the space in the same line.

The phenomenon of language change ..... (**PROBABLE**) attracts more public notice and more ..... (**DISAPPROVE**) than any other linguistic issue.

Part 3: Think of one word only which can be used appropriately in all three sentences.

A day witness can often provide detailed corroboration, thus having a dramatic ..... on the outcome of a complex legal case.

It's generally agreed that the .... of television in the modern world is considerable.

Martha Graham played a major role in developing the theory of modern dance, so extending her ..... to a whole new generation of dancers.

Part 4: Complete the second sentence so that it has a similar meaning to the first sentence, using the word given. Do not change the word given. You must use between three and eight words, including the word given.

Do you mind if I watch you while you paint? **- objection**

Do you ..... you while you paint?

(Cambridge University Press, 2002)

Abbildung 4: CPE – "Use of English" (Auszug)

Grammatik in der "Zentralen Oberstufenprüfung des Goethe-Instituts" (ZOP)

Die Zentrale Oberstufenprüfung des Goethe-Instituts enthält einen Prüfungsteil "Ausdrucksfähigkeit", über den die Kandidaten erfahren:

> In diesem Prüfungsteil sollen Sie zeigen, dass Sie über differenzierte Grammatik- und Wortschatzkenntnisse verfügen. Sie formulieren einen Text um, wobei Sie vorgegebene Ausdrücke verwenden, und lösen Aufgaben zu verschiedenen Aspekten der Grammatik (z.B. Aufgaben zu Präpositionen, Modalverben, Nebensätzen und Partizipialkonstruktionen) und zum Wortschatz (Sie suchen Ausdrücke, die das Gegenteil eines vorgegebenen Wortes bedeuten, und Sie finden neue Formulierungen für Sätze, ohne den Sinn zu verändern), (Goethe-Institut Inter Nationes, 2002).

Im Prüfungsteil "Ausdrucksfähigkeit" sind die Beispiele kontextualisiert und gleichzeitig in Aufgabenstellungen zu je einem Grammatikthema eingebunden. Auszüge aus dem Grammatiktest der ZOP sind in Abbildung 5 abgedruckt.

a) Ergänzen Sie die fehlenden Präpositionen (und wenn nötig den dazugehörenden Artikel):

→ In der Bundesrepublik sorgte in den 70er Jahren ein Film _____ Aufregung, der sich _____ Fall Kaspar Hauser beschäftigte. Die hohen Besucherzahlen bewiesen erneut das Interesse _____ allem, was _____ Kaspar Hauser zusammenhängt. ...

b) Schreiben Sie den folgenden Text neu. Ersetzen Sie dabei die hervorgehobenen Ausdrücke durch die rechts in Klammern angegebenen. Dadurch verändert sich einiges in der Textkonstruktion.

→ Bis heute besteht keine Klarheit (**klar**) über die wirkliche Identität von Kaspar Hauser. (**wer**) ...

c**) Erweitern Sie in den folgenden Sätzen die hervorgehobenen Satzteile zu Satzgefügen:**

Der Amtsarzt legte nach eingehender Untersuchung Kaspar Hausers seinen Bericht vor.

→ Der Amtsarzt legte seinen Bericht vor, ...

d) Verkürzen Sie die folgenden Sätze so, daß sie keinen Nebensatz mehr haben:

Als Kaspar Hauser im Jahre 1833 starb, stand die Welt vor einem Rätsel.

→ ... stand die Welt vor einem Rätsel.

e) Ersetzen Sie die hervorgehobenen Verben durch eine andere Ausdrucksweise ohne Modalverb.

Das Rätsel um Kaspar Hauser konnte nicht gelöst werden.

→ ..., das Rätsel um Kaspar Hauser ...

f) Vervollständigen Sie die Sätze, ohne den Sinn zu verändern:

Unzählige Publikationen sind über das Findelkind geschrieben worden.
→ Das Findelkind ist Thema ...
→ Unzählige Publikationen haben das Findelkind ...
→ Unzählige Publikationen ... von dem Findelkind.
→ In unzähligen Publikationen wurde das Findelkind ...

g) Finden Sie einen Ausdruck, der zu dem hervorgehobenen Wort im Gegensatz steht:

Viele seiner Zeichnungen sind erhalten.
Wären sie ... gäbe es heute noch weniger Material zum Rätsel Kaspar Hauser.

(Goethe-Institut Inter Nationes, ohne Jahr; alte RS im Original)

Abbildung 5: ZOP – "Ausdrucksfähigkeit" (Auszug)

Grammatik im "Kleinen Deutschen Sprachdiplom" (KDS)

Das Kleine Deutsche Sprachdiplom (KDS), welches auch als Nachweis der Sprachkenntnisse für den Hochschulzugang anerkannt wird, enthält einen Prüfungsteil "Aufgaben zur Ausdrucksfähigkeit" (siehe Abbildung 6). Er wird in einer Broschüre für Kandidaten wie folgt beschrieben:

> In diesem Prüfungsteil sollen Sie nachweisen, dass Sie gute Grammatik- und Wortschatzkenntnisse in der deutschen Sprache haben. Die Aufgaben gliedern sich in mehrere Abschnitte zu verschiedenen Grammatik-Kapiteln, z. B. Gebrauch der Präpositionen, der Modalverben, der Nebensätze, Synonyme – Antonyme etc. Sie müssen Einsetzaufgaben lösen und Umformungen vornehmen (Goethe-Institut, 2000: 6).

ZOP und KDS bedienen sich in diesem Prüfungsteil eines ähnlichen Formats.

a) Ergänzen Sie die fehlenden Präpositionen:

Der Handel (**mit**) exotischen Gewürzen.
Schon _____ 6000 Jahren benutzt der Mensch in Europa heimische Gewürze und Kräuter _____ Verfeinerung seiner Speisen. [usw.]

b) Erklären Sie kurz mit eigenen Worten, welche Bedeutung die folgenden Ausdrücke im Text haben:

Was sind "heimische" Gewürze? ...

c) Bilden Sie aus dem hervorgehobenen Satzteil einen Nebensatz!

Gewisse Gewürze nimmt man zur besseren Verträglichkeit von Speisen.
Gewisse Gewürze nimmt man, ...

d) Ersetzen Sie die hervorgehobenen Satzteile durch ein passendes Modalverb (dürfen, können, mögen, müssen, sollen oder wollen), und formen Sie entsprechend um.

Es ist unbedingt notwendig, manche Gewürze zu mahlen.
Manche Gewürze ... unbedingt ...

e) Bilden Sie aus dem Relativsatz jeweils ein Partizipialattribut!

Aus Gebieten, die man neu entdeckt hatte, kamen unbekannte Gewürze in die Alte Welt.
Aus ... kamen unbekannte Gewürze in die Alte Welt.

(Goethe-Institut Inter Nationes, ohne Jahr)

Abbildung 6: KDS – "Aufgaben zur Überprüfung der Ausdrucksfähigkeit" (Auszug)

Grammatik in der "Deutschen Sprachprüfung für den Hochschulzugang" (DSH)

In der "Rahmenordnung über Deutsche Sprachprüfungen für das Studium an deutschen Hochschulen" ist festgelegt, dass die DSH einen Prüfungsteil "Leseverstehen und wissenschaftssprachliche Strukturen" enthält. In der "DSH-Musterprüfungsordnung" wird dazu ausgeführt:

> Die Aufgabenstellung im Bereich Strukturen beinhaltet das Erkennen, Verstehen und Anwenden wissenschaftssprachlich relevanter Strukturen. Diese Aufgabenstellung soll die Besonderheiten des zugrundegelegten Textes zum Gegenstand haben (z. B. syntaktisch, wortbildungs-morphologisch, lexikalisch, idiomatisch, textsortenbezogen) und kann u. a. Ergänzungen, Fragen zum Verstehen komplexer Strukturen sowie verschiedene Arten von Umformungen (Paraphrasierung, Transformation) beinhalten. Sie soll vom Umfang 25 % dieser Teilprüfung umfassen (HRK/KMK, 2004).

Beim DSH-Grammatiktest handelt es sich nach der Musterprüfungsordnung nicht um einen Prüfungsteil, sondern um ein Element eines Prüfungsteils, das separat bewertet wird. Aus Gründen der Prägnanz bleibe ich bei der Bezeichnung DSH-Grammatiktest. Eine detailliertere Beschreibung des DSH-Grammatiktests findet sich im DSH-Handbuch, das vom Fachverband Deutsch als Fremdsprache (FaDaF) für die Ausrichter der DSH erstellt wurde. Hier finden die Ausrichter Hinweise zum Format des Tests und eine Musterprüfung. Diese beziehen sich auf die alte Rahmenordnung. Abgesehen von einer Anbindung an das Leseverstehen und einer Verringerung des Umfangs sind jedoch keine Änderungen des Formats geplant, so dass ich mich auf die Erläuterungen im DSH-Handbuch stützen werde. Damit wird ein Ausschnitt der Prüfungswirklichkeit wiedergegeben, denn die Rahmenordnung bzw. die Aussagen im DSH-Handbuch werden unterschiedlich ausgelegt: Betrachtet man die Musterprüfungen, welche von Studienkollegs und Sprachlehrinstituten veröffentlicht werden, trifft man auf eine bunte Prüfungsvielfalt.

Im DSH-Handbuch heißt es zum DSH-Grammatiktest:

> Der Prüfungsteil 'wissenschaftssprachliche Strukturen' befasst sich sowohl rezeptiv als auch produktiv mit wissenschaftssprachlich relevanten Strukturen in einem vorgegebenen Text. Er will weder die Kenntnis grammatischer Terminologie erfragen, noch theoretisches Wissen der Kandidatinnen und Kandidaten überprüfen, z. B. Kenntnisse über die Bildung des Passivs. Es soll vielmehr geprüft werden, ob die [...] Strukturen kontextgebunden erkannt, verstanden und angewendet werden können [...] (FaDaF, 2001: 7/2).

Im "DSH-Handbuch für Prüferinnen und Prüfer" werden zur Konzeption dieses Prüfungsteils folgende Hinweise gegeben: Insgesamt elf sprachliche Phänomene stellen den "Kanon" dar. Kein Strukturproblem soll mehr als zwei Mal behandelt werden. Als

Ausgangsmaterial dient ein zusammenhängender Text. Die Aufgabenstellung lautet: "Füllen Sie die Lücken aus, ohne die Textinformation zu verändern!" (FaDaF, 2001: 7/2).

Hintergrund dieser eindeutigen Aussage dürfte die Beobachtung sein, dass häufig linguistische Fachbegriffe in den Aufgabenstellungen verwendet werden, was der Zielsetzung "grammatische Terminologie wird nicht erfragt" jedoch widerspricht. Andererseits stellt diese Empfehlung eine deutliche Einschränkung der (alten oder neuen) DSH-Rahmenordnung (HRK, 2000; HRK/KMK, 2004) dar, welche Aufgabenstellungen in Form von "Ergänzungen, Fragen zum Verstehen komplexer Strukturen sowie verschiedene Arten von Umformungen (Paraphrasierung, Transformation)" vorsieht. Wie dem auch sei: Mit der Aufgabenstellung kann eine Reihe von grammatischen Phänomenen erfasst werden. Sie wird im Handbuch anhand eines Texts zum Thema "Flurbereinigung" erläutert (siehe Abbildung 7). Da anders als im KDS oder in der ZOP nicht vorgegeben wird, welche Transformation von den Testteilnehmern verlangt wird, muss zunächst die Struktur erkannt werden, dann erst kann die Lücke ausgefüllt werden. Mit den Beispielen, die im DSH-Handbuch folgen, wird unterstrichen, dass der DSH-Grammatiktest kein Sprachwissenstest sein soll. Die Autoren sehen einen kontextualisierten Sprachverwendungstest vor.

Der DSH-Grammatiktest kommt laut "DSH-Handbuch für Prüferinnen und Prüfer" (FaDaF, 2001) mit einem Aufgabentyp aus. Die Items bestehen aus sprachlichen Transformationen, die in einen Textzusammenhang eingebettet sind. Durch die Beschränkung auf einen Aufgabentyp unterscheidet sich der DSH-Grammatiktest von anderen Tests der Grammatik. Die Beschränkung dürfte gewählt worden sein, damit er leichter zu erstellen ist; das ist bei nicht standardisierten Tests wichtig, bei zentral gestellten Tests nicht. In der Praxis gibt es freilich auch andere Formate. Welche Auswirkungen unterschiedliche Aufgabentypen haben, ist Thema einer Studie, die in Kapitel 4.1 vorgestellt wird.

Der DSH-Grammatiktest unterscheidet sich deutlich vom TOEFL-Prüfungsteil "*Structure and Written Expression*", der durchgehend im Multiple-Choice Format konzipiert ist. Bei einigen Items des CPE-Prüfungsteils "*Use of English*" geht es um Transformationen, welche mit denen aus dem DSH-Grammatiktest vergleichbar sind. "*Use of English*" ist jedoch umfangreicher. Deutliche Parallelen bestehen zwischen den

jeweiligen Prüfungsteilen aus dem KDS bzw. der ZOP und dem DSH-Grammatiktest. Im Unterschied zum KDS und zur ZOP sollen die Items laut DSH-Handbuch in der DSH nicht nach Strukturen geordnet werden und den Kandidaten wird nicht mitgeteilt, welche Struktur zu verwenden ist.

Füllen Sie die Lücken aus, ohne die Textinformation zu verändern! Die Unterstreichungen sollen Ihnen bei der Lösung helfen:	
Fragt man nach den Ursachen der in Deutschland in den letzten Jahren verstärkt auftretenden Überschwemmungen,	Fragt man, *warum in Deutschland in den letzten Jahren verstärkt Überschwemmungen auftreten*,
so findet man einen wesentlichen Faktor in der Flurbereinigung, die in den siebziger und achtziger Jahren im Zuge der EG-Agrarpolitik vorgenommen wurde.	so findet man einen wesentlichen Faktor in der ....Flurbereinigung.
Die Flurbereinigung ist ein Eingriff in die Landschaft, der ihre natürliche Beschaffenheit nachhaltig verändert.	Die Flurbereinigung ist ein Eingriff in die Landschaft, der ....führt.
Ursprünglich lag der Flurbereinigung eine sehr sinnvolle Idee zugrunde: die durch Jahrhunderte lange Erbteilungen entstandenen kleinen und verstreuten Felder sollten zusammengelegt und flächenmäßig unter den Besitzern neu aufgeteilt werden.	Ursprünglich lag der Flurbereinigung eine sehr sinnvolle Idee zugrunde: Die kleinen und zerstreuten Felder, ... sollten zusammengelegt und flächenmäßig unter den Besitzern neu aufgeteilt werden.
[...]	

(FaDaF, 2001: 7/4; vollständiger Test: Anhang 1, Seite 354)

Abbildung 7: DSH-Grammatiktest "Flurbereinigung" – Prototyp aus dem DSH-Handbuch

3.2. Fragen an den DSH-Grammatiktest

Übersicht: Kapitel 3.2
In diesem Kapitel richte ich Fragen an den DSH-Grammatiktest, die sich aus der Sprachwissenschaft und der Testmethodik ergeben: Handelt es sich um einen Test der expliziten oder der impliziten Grammatik? Werden Sprachentwicklungsstufen in dem Grammatiktest berücksichtigt? Handelt es sich um einen Kompetenztest oder um einen Performanztest? Handelt es sich um einen direkten oder um einen indirekten Test? Handelt es sich um einen normorientierten oder um kriteriumsorientierten Test? Welche Funktion hat der Test? Aussagen zur Qualität des DSH-Grammatiktests werden anhand der Kriterien zur Nützlichkeit von Bachman und Palmer erhoben. Mit diesen Fragen erfolgt eine Analyse des DSH-Grammatiktests, es werden Stärken und Schwächen herausgearbeitet und offene Fragen identifiziert.

Im folgenden richte ich Fragen an den DSH-Grammatiktest, anhand derer die Besonderheiten des Prüfungsteils herausgearbeitet werden sollen. Damit verfolge ich folgende Ziele: die Einordnung des DSH-Grammatiktests in sprachwissenschaftliche und testtheoretische Konzepte, Identifizierung offener Fragen sowie Erhebung der Nützlichkeit. Zur Nützlichkeit beziehe ich mich auf die in Kapitel 2.2 vorgestellten Kriterien von Bachman und Palmer. Weitere Fragen, die an Grammatiktests gestellt werden sollten, sind: Wird explizites oder implizites Grammatikwissen geprüft? Werden Sprachentwicklungsstufen berücksichtigt? Handelt es sich um einen Kompetenztest oder um einen Performanztest, um einen direkten oder um einen indirekten Test? Ist der DSH-Grammatiktest normorientiert oder kriteriumsorientiert? Wie ist die Testfunktion?

Explizites oder implizites Grammatikwissen?

Mit einer exakten Definition der explizit/implizit-Unterscheidung ringt die angewandte Sprachwissenschaft seit langem (Raupach, 2002). Die folgende Definition von Bialystok kann als Grundlage für diese ungenaue Unterscheidung gelten:

> Explicit Linguistic Knowledge contains all the conscious facts the learner has about the language and the criterion for admission to this category is the ability to articulate those facts [...]. Implicit Language Knowledge is the intuitive information upon which the language learner operates in order to produce responses in the target language. Whatever information is automatic and is used spontaneously in language tasks is represented in Implicit Linguistic Knowledge (Bialystok, 1978; zit. n. Raupach, 2002: 104).

Auch nach Ellis liegt der Unterschied zwischen explizitem und implizitem Sprachwissen in dem Grad der Bewusstheit. Explizites Wissen kann eine Kenntnis der Metasprache einbeziehen, es bezieht sich jedenfalls auf eine Bewusstheit im Umgang mit sprachlichen Phänomenen. Auf explizites Sprachwissen hat man nur einen verlangsamten Zugang, auf implizites Sprachwissen kann man rasch zugreifen (vgl. Ellis, 2001: 252). Verdeutlichen lässt sich der Unterschied am Unterricht: DeKeyser spricht von einer expliziten Grammatik*vermittlung*, wenn Regeln erklärt werden oder wenn die Aufmerksamkeit der Lerner auf ein bestimmtes Phänomen aus der Grammatik gerichtet wird. Implizit geht ein Unterricht vor, in dem keine Regeln formuliert werden und die Aufmerksamkeit der Lerner nicht auf bestimmte Phänomene gerichtet werden (DeKeyser, 1995).

Mit Blick auf Grammatiktests wirft der implizit/explizit-Unterschied wichtige Fragen auf: Welches Grammatikwissen sollte in Sprachtests für den Hochschulzugang geprüft werden? Wird im DSH-Grammatiktest explizites oder implizites Grammatikwissen geprüft? Wie explizit muss Grammatik zur Vorbereitung auf den DSH-Grammatiktest vermittelt werden?

- Zur ersten Frage: Ellis argumentiert, dass implizites Wissen als Indikator dafür gelten kann, ob der Erwerb einer Struktur stattgefunden hat, und dass daher implizites Wissen geprüft werden sollte (Ellis, 2001: 253). Diese Argumentation trifft auch auf Sprachtests für den Hochschulzugang zu: Es sollte implizites Wissen geprüft werden.

- Wie verhält es sich nun beim DSH-Grammatiktest? Es wurde bereits eine Anweisung aus dem DSH-Handbuch zitiert, nach der die Strukturen "kontextgebunden erkannt, verstanden und angewendet werden" sollen (FaDaF, 2001: 7/2). Erkennen und Ver-

stehen sind Aspekte, die ich eher einem expliziten Vorgehen zurechnen würde. Bewertet wird jedoch allein das Ergebnis, die Anwendung – möglicherweise reicht implizites Wissen zur Anwendung aus. Anders als in der ZOP oder im KDS werden im DSH-Grammatiktest keine metasprachlichen Hinweise zur Bearbeitung der Aufgaben gegeben. Es hilft den Testteilnehmern beim DSH-Grammatiktest also nicht, wenn sie mit der linguistischen Metasprache vertraut sind. Verwendung und Vermittlung von Metasprache führen zu explizitem Wissen, der Verzicht darauf deutet auf implizites Grammatikwissen. Andererseits erfolgt die Bearbeitung des Grammatiktests nicht unter Zeitdruck. Die Kandidaten haben also Zeit, sich mit den Phänomenen auseinanderzusetzen und können auf explizites Wissen zurückgreifen. Dies deutet darauf hin, dass auch explizites Grammatikwissen erfasst wird. Es wird deutlich, dass es sich um Kategorien handelt, die man im konkreten Fall nur schwer zuordnen kann. Vor allem der letztgenannte Aspekt, die intensive Beschäftigung mit einem Satz, ist für mich ein entscheidender Hinweis darauf, dass explizites Grammatikwissen eingesetzt wird.

- Was die Vorbereitung auf den DSH-Grammatiktest betrifft, könnte sie sowohl explizit als auch implizit verlaufen. In der Praxis geht es dabei weniger um die implizit/explizit Unterscheidung als vielmehr um den Grad der Explizitheit. Diesem Gedanken gehe ich im Abschnitt "Nützlichkeit" unter "Auswirkungen" nach (siehe Seite 90).

Berücksichtigung von Sprachentwicklungsstufen?

Pienemann konnte in Studien zeigen, dass der Spracherwerb in Entwicklungsstufen verläuft, die einen hohen Grad an Allgemeingültigkeit besitzen (Pienemann, 1984; 1989; 1998; siehe auch Ellis, 2001; Multhaup, 2002). Aus diesen Erkenntnissen leitete er die *Teachability Hypothesis* und als Weiterentwicklung die *Processability Theory* ab. Der Verdienst dieser Theorien liegt u. a. darin, dass begründete Aussagen gemacht werden können, zu welchem Zeitpunkt des Spracherwerbs Lerner in der Lage sind, bestimmte sprachliche Strukturen zu erlernen und produktiv zu nutzen. Auch für Sprachtests können die Sprachentwicklungsstufen genutzt werden: Es ist prinzipiell möglich, Grammatiktests auf bestimmte Sprachentwicklungsstufen abzustimmen. Die

Ergebnisse dieser Tests könnten im Sinne eines Diagnosetests Informationen über die Sprachentwicklungsstufen der Testteilnehmer bieten (Ellis, 2001; Spada/Lightbown, 1993).

Dies ist beim DSH-Grammatiktest jedoch nicht der Fall. Sprachentwicklungsstufen werden allenfalls zufällig oder indirekt berücksichtigt. Man könnte bei Antworten, die von der erwarteten Norm abweichen, die Frage stellen, welche Kandidaten jeweils über fortgeschrittenere Sprachkenntnisse verfügen und dies bei der Bewertung berücksichtigen. Ein Beispiel aus dem Prototyp-Grammatiktest aus dem DSH-Handbuch: Den präpositionalen Ausdruck "Bei der Durchführung der Flurbereinigung …" sollen die Kandidaten in einen Nebensatz umformen, also beispielsweise: "Als die Flurbereinigung durchgeführt wurde…" Typische Antworten waren:

1) "Als der Flurbereinigung durchgeführt wurde"
2) "Weil die Flurbereinigung durchgeführt wurde"
3) "Wenn man die Flurbereinigung durchgeführt wurde"
4) "Die Flurbereinigung durchgeführt wurde"
5) "Flurbereinigung durchgeführt wurde"

Hier könnte man nun fragen, ob die Wahl einer unpassenden Konjunktion ("weil" statt "als" in Antwort 2) auf eine niedrigere Spracherwerbsstufe verweist als der falsche Artikel ("der" statt "die" in Antwort 1). Antwort 3 verdeutlicht Unsicherheiten im Gebrauch des Passivs, bei Antwort 4 fehlt die Konjunktion, bei Antwort 5 fehlen Konjunktion und Artikel. Falls eine Zuordnung zu unterschiedlichen Spracherwerbsstufen gelingt, könnte dies bei der Auswertung berücksichtigt werden. Möglicherweise würde man sogar zu einer anderen Bewertung gelangen, denn die Spracherwerbsstufen verlaufen nicht analog zum Schwierigkeitsgrad der Phänomene.

An diesem Beispiel werden auch Schwierigkeiten deutlich. Die geforderte Umformung von präpositionalen Ausdrücken in Nebensätze ist kaum einer der bei Pienemann beschriebenen Spracherwerbsstufen zuzuordnen, da sie eine Vertrautheit mit allen Stufen voraussetzt. Das trifft auch auf die übrigen sprachlichen Phänomene zu, die im Grammatiktest behandelt werden. Laut DSH-Handbuch können folgende Strukturen im DSH-Grammatiktest geprüft werden:

Die folgenden Strukturen haben sich als relevant erwiesen und bilden den Aufgabenkatalog:
- Attributionen
- Infinitivstrukturen (Infinitive/undeklinierte Partizipien)
- Relativsätze
- Präpositionalphrasen/Nebensätze (insbesondere in Bezug auf logische Beziehungen und Textstrukturen)
- Direkte Rede/Indirekte Rede
- Aktiv/Passiv/Passiversatz
- deiktische Beziehungen
- logische Junktoren/textsteuernde Elemente
- Funktionsverbgefüge
- Nominalisierungen/Verbalisierungen
- Hypothesen/Konjunktive (FaDaF, 2001: 7/2).

Der DSH-Grammatiktest zielt nicht darauf ab, Spracherwerbsstufen abzubilden.

Kompetenztest oder Performanztest, direkter oder indirekter Test?

Der DSH-Grammatiktest ist ein Kompetenztest. Seine Authentizität ist gering, eine konkrete Anwendungssituation ist kaum vorstellbar (siehe auch "Nützlichkeit: Authentizität und Interaktivität", Seite 90). In sprachlichen Performanztests stellt man die Frage, ob die Kandidaten ihre Sprachfähigkeit in einer spezifischen Situation anwenden können. Man bedient sich in der Regel einer möglichst realitätsnahen Sprachverwendungssituation. Sprachliche Kompetenztests haben die allgemeine Beherrschung von Sprache unabhängig von einer konkreten Anwendungssituation zum Gegenstand. Sie bestehen aus dem Abfragen von Einzelfertigkeiten. In der Regel lassen sich Kompetenztests eher durch einen Lerneinsatz verbessern als Performanztests. Sie werden gerne als Lernfortschrittstests oder Kursabschlussprüfungen eingesetzt, die auf den im Unterricht behandelten Themen beruhen (siehe Kapitel 2.1, Seite 29 ff).

Bei einem direkten Testverfahren kann man ohne weitere Interpretation von der Testleistung auf die Leistung in der Wirklichkeit schließen. Je indirekter ein Testverfahren ist, desto größer wird die Notwendigkeit zur Interpretation der Testergebnisse. Vorteile direkter Tests sind eine möglicherweise hohe Validität: Durch die Nachbildung der realen Sprachverwendungssituation bekommt man das Problem der Konstruktvalidität in den Griff (siehe Kapitel 2.1, Seite 49 ff). Aus der Beschreibung der sprachlichen Anforderungen, mit der die reale Sprachverwendungssituation zu meistern ist, kann eine mehr oder weniger konkrete Zielvorgabe für den Test abgeleitet werden. Die Schwierigkeiten direkter Performanztests liegen in der Übertragbarkeit der Ergebnisse auf

andere Anwendungssituationen sowie in der Bewertung der Leistungen (Fulcher, 2000; McNamara, 1997).

Indirekte Kompetenztests erlauben in der Regel eine objektive Bewertung und eine einfache Verwaltung. Die Interpretation der Ergebnisse ist schwieriger, sie lässt sich im Prinzip aber auf verschiedene Sprachverwendungssituationen übertragen. Die Inhaltsvalidität (erfasst der Inhalt der Test-Items das Konstrukt in seinen wichtigsten Aspekten?) indirekter Tests ist wegen des großen Interpretationsspielraumes gering.

Die auf den ersten Blick so einfach erscheinende Frage, ob der DSH-Grammatiktest direkt oder indirekt vorgeht, ist nicht auf Anhieb zu klären. Die Unsicherheiten bei der Bestimmung des Testkonstrukts erschweren die Zuordnung. Wenn das Konstrukt des DSH-Grammatiktests allein "Erkennen, Verstehen und Anwenden von wissenschaftssprachlich relevanten Strukturen" lautet, kann dem Test eine direkte Vorgehensweise nicht abgesprochen werden. Ob die Kandidaten in der Lage sind, sprachlich komplexe Transformationen vorzunehmen, kann allerdings nicht das alleinige Anliegen eines DSH-Prüfungsteils sein. Da weder von ausländischen noch von deutschen Studierenden derartige Transformationsaufgaben im Studium verlangt werden, lässt sich das Testkonstrukt nur legitimieren, wenn die im DSH-Grammatiktest gezeigten Fertigkeiten den Studierenden bei Sprachverwendungssituationen im Maschinenbau- oder Medizinstudium ebenfalls helfen. Nur wenn sie eine Voraussetzung für das Verständnis von Texten oder Vorlesungen, das Formulieren in Klausuren oder die Teilnahme an Klausuren sind oder wenn sie ein Indikator für gute Leistungen in derartigen Sprachverwendungssituationen darstellen, ist es legitim, von ausländischen Studienbewerbern sprachliche Aktivitäten zu verlangen, auf die sie im Studium nicht treffen. Vor diesem Hintergrund ist der DSH-Grammatiktest als indirekter Kompetenztest einzuordnen.

Bezugsgröße: Normorientierter oder kriteriumsorientierter Test?

Sprachtests für den Hochschulzugang sind von ihrer Zielsetzung her kriteriumsorientiert (und nicht normorientiert; siehe Kapitel 2.1, Seite 25): Sie sollen ermitteln, ob die Kandidaten "ausreichende Sprachkenntnisse für das Studium" erreicht haben. Anliegen ist nicht der Vergleich individueller Ergebnisse mit den Ergebnissen anderer wie bei normorientierten Tests. Auch der DSH-Grammatiktest ist im Prinzip kriteriumsorientiert, allerdings ist sein genauer Beitrag zum Kriterium ungeklärt. Wie können Schwellenwerte und Äquidistanzen ermittelt werden, wenn der Test selbst wenig Hinweise dafür gibt? Auch die anderen Grammatiktests werfen dieses Problem auf, denn sie sind indirekte Kompetenztests, die wenig Information über das Erreichen des Kriteriums bieten. Wie sollen Schwellenwert und Äquidistanzen beim DSH-Grammatiktest festgelegt werden? Es wurde in Kapitel 2.1 (Seite 25 f) bereits darauf hingewiesen, dass in der DSH-Rahmenordnung zu einer fragwürdigen Abkürzung geraten wird: Demnach liegen die Schwellenwerte bei 57, 67 und 82 Prozent (HRK/KMK, 2004: § 5 der DSH-Musterprüfungsordnung), wobei das Testkriterium nur oberflächlich beschrieben wird (auf der Rückseite des Zeugnisses). Ich gehe davon aus, dass sich die durchführenden Institute entweder starr an diese Vorschrift halten oder dass die Testleistungen mit den Leistungen im Unterricht verglichen werden. Der DSH-Grammatiktest ist daher nicht wirklich kriteriumsorientiert, er wird vielmehr an einer nicht näher bestimmten, erfahrungsorientierten Norm orientiert.

Funktion: Kursabschlussprüfung oder Feststellungsprüfung?

Zentrale Feststellungsprüfungen stellen hohe Ansprüche an die Reliabilität und die Validität (Kapitel 2.1, Seite 13 ff). Beim DSH-Grammatiktest sind diese beiden Aspekte in der Praxis nicht von zentraler Bedeutung. Hier stehen der Bezug zum Unterricht, die Nähe zu den Prüfungskandidaten und die Testökonomie im Vordergrund. Es offenbart sich eine Testtradition, welche von Lehrenden, von der Institution, von persönlichen Einschätzungen ausgeht. Man kann von einem Bezug zum Unterricht ausgehen. Man kann davon ausgehen, dass es eine Nähe zu den Prüfungskandidaten gibt. Dadurch wird der Einsatz des DSH-Grammatiktests als Kursabschlussprüfung deutlich. Das ist ungewöhnlich für eine Sprachprüfung für den Hochschulzugang.

Nützlichkeit: Wie nützlich ist der DSH-Grammatiktest?

Kriterien zur Nützlichkeit von Sprachtests sowie die Bestandteile der Nützlichkeit nach Bachman und Palmer – Reliabilität, Konstruktvalidität, Authentizität, Interaktivität, Auswirkungen und Ökonomie – wurden in Kapitel 2.2 (Seite 49 ff) vorgestellt. Nun geht es um die Nützlichkeit des DSH-Grammatiktests.

Reliabilität: Wie reliabel ist der DSH-Grammatiktest? Es gibt keinen Grund für die Annahme, dass die interne Konsistenz des DSH-Grammatiktests gering ist. Der Schwierigkeitsgrad der Items dürfte sich zwar unterscheiden, weil unterschiedliche sprachliche Phänomene geprüft werden. Das Testkonstrukt einzelner Items, die produktive Grammatikkompetenz, dürfte sich jedoch ähneln. Schon weniger sicher bin ich bei der Frage nach der Testwiederholungsreliabilität: Es ist durchaus möglich, dass ein deutlicher Übungseffekt auftreten kann. Doch auch bei der Testwiederholungsreliabilität dürfte kein besonders ausgeprägtes Problem liegen. Dies besteht meiner Ansicht nach vor allem bei der Paralleltestreliabilität: Wie auch bei den anderen Prüfungsteilen der DSH kann auch beim DSH-Grammatiktest nicht von einer hohen Paralleltestreliabilität ausgegangen werden. Es dürfte deutliche Unterschiede zwischen Versionen an einem Institut und vor allem an unterschiedlichen Instituten geben (siehe auch Kapitel 2.2, Seite 43). Da man nicht einmal davon ausgehen kann, dass sich die Testmethoden-Merkmale der DSH-Grammatiktests gleichen, müsste man untersuchen, welche Auswirkungen unterschiedliche Merkmale auf die Reliabilität und die Validität haben.

Konstruktvalidität: Welche Interpretationen, welche Schlussfolgerungen ermöglicht ein Grammatiktest mit Blick auf das Testkonstrukt von Sprachtests für den Hochschulzugang? Rea-Dickins (1997: 95) stellt mit Blick auf Grammatiktests fest: *We have [...] incomplete knowledge about the nature of the construct*. Diese Aussage trifft nach wie vor zu, die Frage nach der Konstruktvalidität von Grammatiktests kann nicht zufrieden stellend beantwortet werden. Zur Bestimmung des Konstrukts von Grammatiktests gibt es zwei Ansätze: einen theoriegeleiteten und einen empirischen.

Die *theoriegeleitete Bestimmung*, auf die ich zunächst eingehen möchte, geht von Annahmen über die Elemente der Sprachkompetenz und der Grammatikkompetenz aus. Am Anfang stehen also die Fragen, was der Gegenstand von Grammatiktests ist und

was unter Grammatikkompetenz zu verstehen ist. Der Begriff "Grammatik" ist mehrdeutig (vgl. Bußmann, 2002: 259-260):

- Erstens ist die Grammatik als Lehrbuch zu nennen: In einer Grammatik wird das Regelsystem umfassend als linguistische Grammatik oder passend für eine bestimmte Zielgruppe als didaktische oder pädagogische Grammatik beschrieben (Götze, 2001a).

- Wenn zweitens die Grammatik als Sprachlehre gemeint ist, so geht es im eigentlichen Wortsinn um die Lehre von den Buchstaben, im weiteren Sinne um die Lehre von der Schrift. Laut Dudenredaktion versteht man unter Grammatik als Sprachlehre den "Teil der Sprachwissenschaft, der sich mit den sprachlichen Formen und deren Funktion im Satz, mit den Gesetzmäßigkeiten, dem Bau einer Sprache beschäftigt" (Dudenredaktion, 2001). In der Sprachwissenschaft zählt man auf der Satzebene die Lautlehre (Phonetik), die Formenlehre (Morphologie) und die Satzlehre (Syntax) zur Grammatik und weitere Elemente auf der Gesprächs- bzw. Textebene (siehe die Erläuterungen unten zu Purpura, 2004).

- Die dritte Bedeutung kann durch die Unterscheidung zwischen *langue* und *parole* von de Saussure erfasst werden: Mit Grammatik bezeichnet man hier das Sprachsystem (*langue*). Es handelt sich um das Regelsystem, das der Sprache zugrunde liegt. Es lässt sich nur indirekt über die sprachlichen Äußerungen (*parole*) erschließen.

- Ein verwandter Ansatz ist viertens die Grammatik als theoretisches Modell zur Abbildung der Sprachkompetenz. Die generative Transformationsgrammatik von Chomsky (1965: 3-9) beruht auf dem letztgenannten Ansatz.

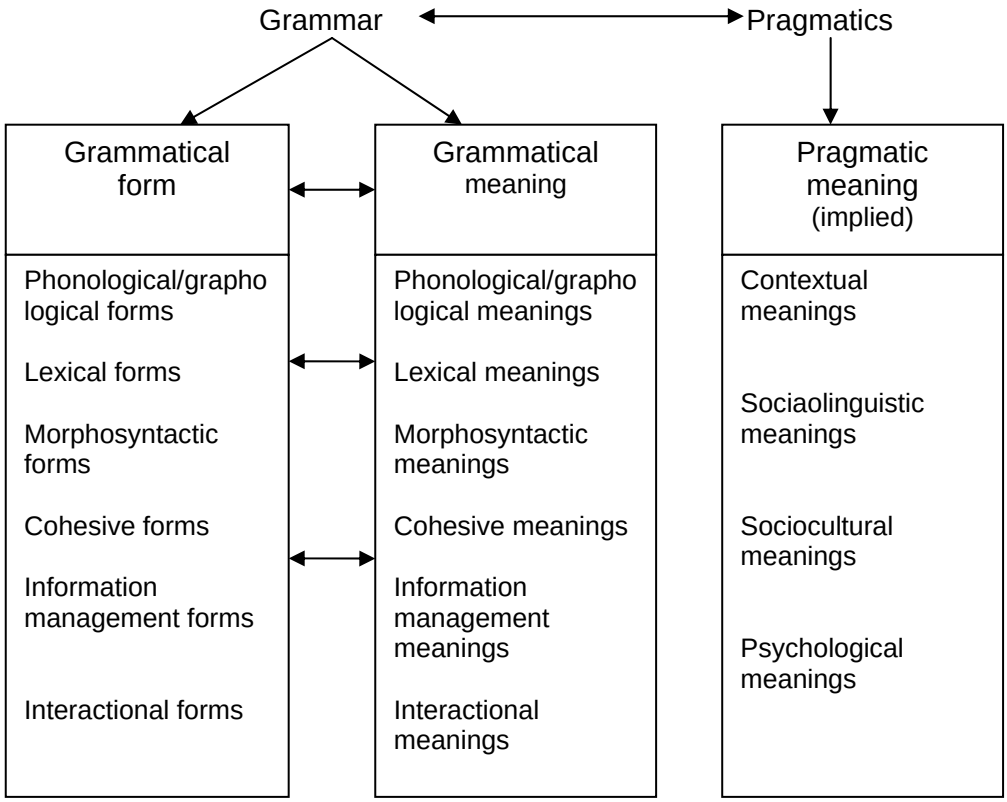
Für die Interpretation von Grammatiktests ist die Unterscheidung zwischen der Kompetenz des Sprechers und dem Regelsystem relevant. Dies Unterscheidung hebt Helbig hervor:

> - eine Grammatik A: das der Sprache selbst innewohnende Regelsystem, unabhängig von dessen Beschreibung durch die Linguisten und von dessen Beherrschung durch die Sprecher;
> - eine Grammatik B: die Abbildung des der Sprache selbst innewohnenden Regelsystems durch die Linguistik;
> - eine Grammatik C: das dem Sprecher interiorisierte Regelsystem (seine "subjektive Grammatik"), auf Grund dessen dieser die betreffende Sprache beherrscht (Helbig, 1993: 21).

Mit einem Grammatiktest zielt man in erster Linie darauf ab, die subjektive Grammatik der Testteilnehmer zu erfassen (also Grammatik C). Dazu werden die von den Kandidaten realisierten sprachlichen Äußerungen mit den erwarteten verglichen. Die Art und Weise, wie dies geschieht, offenbart ein bestimmtes Verständnis der Grammatik, auch eine Einstellung dazu, wie Grammatik (als Sprachlehre) im Unterricht behandelt werden soll.

Wie äußert sich Kompetenz in Grammatik? Purpura (2004) unterscheidet zwischen Grammatikwissen (*grammatical knowledge*) und Grammatikkompetenz (*grammatical ability*). Zum Grammatikwissen zählt Purpura im Sinne der linguistischen Beschreibung Elemente der Phonetik, Lexik und Morphologie auf der Satzebene und auf der Text- bzw. Gesprächsebene Elemente wie Kohäsion, Prosodie oder Gesprächstechnik. Das Grammatikwissen bezieht sich jeweils auf die Kenntnis der grammatischen Form und auf die Kenntnis ihrer Bedeutung. Die Ebene der Pragmatik ist dabei ein verwandter, aber von der Gammatikkompetenz zu trennender Bestandteil (siehe Abbildung 8, Seite 85). Die Grammatikkompetenz setzt sich aus Grammatikwissen und strategischer Sprachkompetenz zusammen. Dementsprechend definiert Purpura auch das Konstrukt von Grammatiktests: Es setzt sich zusammen aus den Bestandteilen der Grammatikkompetenz, welche abhängig von der Aufgabenstellung und dem Kontext unterschiedlich deutlich erfasst werden.

Ergebnisse im Grammatiktest lassen demnach Schlussfolgerungen darüber zu, ob bzw. zu welchem Grad die formale Seite der Sprache und ihre Bedeutungen beherrscht werden, ob produktive Grammatikkompetenz vorhanden ist. Es gibt keinen Grund zur Annahme, dass ein DSH-Grammatiktest, der nach dem Muster im DSH-Handbuch entworfen wurde, mit Blick auf dieses Testkonstrukt entscheidend abweicht. Man könnte allenfalls argumentieren, dass die Bedeutung der sprachlichen Form bei Transformationsaufgaben in den Hintergrund gerät, vor allem mit der Aufgabenstellung "Füllen Sie die Lücken aus, ohne die Textinformation zu verändern!" Schließlich haben Sätze im Aktiv beispielsweise eine etwas andere Bedeutung als Sätze im Passiv. Abgesehen von diesem Einwand gehe ich davon aus, dass sich die Transformationsaufgaben im DSH-Grammatiktest eignen, um das Testkonstrukt "produktive Grammatikkompetenz" zu erfassen.



(nach Purpura, 2004)

Abbildung 8: Grammatikwissen und pragmatisches Wissen

Die schwierigeren Fragen, die gestellt werden müssen, sind: Sind Informationen über die Grammatikkompetenz für den Nachweis ausreichender Deutschkenntnisse so relevant, dass sie in einem eigenen Prüfungsteil bzw. in besonderen Items erhoben werden müssen? Und: Welche weiteren Schlussfolgerungen, welche weiteren Erkenntnisse über die Sprachkompetenz ermöglichen die Leistungen im DSH-Grammatiktest? Die erste Frage (nach der Relevanz) ist nur mit Spekulationen zu beantworten, ich komme am Schluss des Kapitels 4 darauf zurück. Die zweite Frage (Validität) wird bisweilen sehr unterschiedlich beantwortet. Eine häufige Annahme ist, dass Grammatiktests Interpretationen mit Blick auf Konstrukte wie "globale Sprachkompetenz" ermöglichen:

> Grammar is far more powerful in terms of generalisability than any other language feature. Therefore grammar may still be the most salient feature to test (Davies, 1982: 151).

Auch Clapham (1996) setzte Grammatiktests ein, um die globale Sprachkompetenz der Testteilnehmer in ihrer Studie zu erfassen (siehe Kapitel 5.2). Grammatiktests, welche zum Test "globaler Sprachkompetenz" eingesetzt werden, werfen Fragen nach den zugrunde liegenden Bestandteilen der menschlichen Sprachkompetenz auf. In Kapitel 2.1 (Seite 35 f.) wurde bereits darauf hingewiesen, dass sich die Vorstellung Ollers (1979) von der Eindimensionalität der Sprachkompetenz nicht durchsetzte. Bachman und Palmer (1996) argumentieren demgegenüber, dass der Sprachkompetenz ein mehrdimensionales Konstrukt zugrunde liegt. Wie oben dargestellt wurde, argumentiert auch Purpura (2004) in diesem Sinne. Dieser Argumentation folge ich in dieser Arbeit und untersuche in Kapitel 4.2 empirisch, welche Aspekte des mehrdimensionalen Konstrukts vom DSH-Grammatiktest erfasst werden.

Informationen zum Testkonstrukt können anhand *empirischer Befunde* erhoben werden. Ergebnisse in Grammatiktests müssten hoch mit den Ergebnissen in Tests korrelieren, von denen bekannt ist, dass sie globale Sprachkompetenz abbilden. Dies wäre zwar noch kein Nachweis für das Testkonstrukt, wohl aber eine notwendige Bedingung für eine derartige Interpretation.

Bislang existieren nur wenige Studien, in denen der Versuch unternommen wurde, das Testkonstrukt von Grammatiktests näher einzugrenzen. In einigen Studien wurde eine Nähe des Konstrukts von Grammatiktests zum Leseverstehen nachgewiesen. Auch in inhaltlichen Analysen wird Grammatikkompetenz als wichtiger Bestandteil für die Lesekompetenz interpretiert (Bernhardt, 1999: 4). Bei Untersuchungen zu einer Pilotversion des IELTS-Tests beobachtete man eine hohe Korrelation zwischen einem Grammatiktest mit dem IELTS-Prüfungsteil Leseverstehen. Überraschend war, dass die Korrelationen zwischen dem Grammatiktest und vier Leseverstehenstest aus dem IELTS (abgesehen von einer Ausnahme) höher waren als die Korrelation der Ergebnisse in den Leseverstehenstests untereinander. Alderson zog – vorsichtige – Folgerungen in Bezug auf das Konstrukt von Grammatiktests:

> Although we have gathered powerful statistical evidence on the functioning of the tests, we do not have a clear picture of what they are actually testing. [...] In the grammar test they [students] may consciously process and reflect on grammar, whereas in the reading tests they may focus on meaning an information and process the grammar, if at all, only on a subconscious level. The reading tests may tap an automatized grammatical ability, whereas the Grammar test might call upon a reflective

> awareness of grammar. [...] It must be the case that, in some intuitive sense, a reader must process the grammar in a text in order to understand it (Alderson, 1993: 217).

Diese Beobachtung konnte in einer Folgeuntersuchung präzisiert werden: Testteilnehmer, die im Grammatiktest ein hohes Ergebnis erzielten, erreichten auch im Prüfungsteil Leseverstehen ein hohes Ergebnis und konnten dabei besser mit Fachtexten umgehen, selbst wenn diese nicht aus ihrer Disziplin stammten (Clapham, 1996). Zur Übertragbarkeit der Studien auf das Deutsche meint Grotjahn:

> Im Fall einer flektierenden Sprache wie des Deutschen dürften Grammatikkenntnisse im Übrigen vermutlich noch stärker als im Englischen zur Leseleistung beitragen (Grotjahn, 2000b: 20).

Zur Schreibkompetenz sind die Befunde weniger eindeutig: Untersuchungen zum TOEFL Prüfungsteil "Structure and Written Expression", in denen er mit Tests zur Schreibfertigkeit verglichen wurde, führten zu unterschiedlichen Ergebnissen. Die zunächst hohen Korrelationen konnten in anderen Studien nicht wiederholt werden (DeMauro, 1992; Stansfield, 1986). Da Sprachtests bei der Bewertung des Schreibens Grammatik als Bewertungskriterium ausdrücklich erwähnen, ist meiner Ansicht nach eine Ähnlichkeit der Konstrukte von Grammatiktests und Tests zur Schreibfertigkeit durchaus zu erwarten, eher als eine Ähnlichkeit der Konstrukte von Grammatiktests und Tests zum Leseverstehen.

Insgesamt muss man feststellen, dass zur Konstruktvalidität von Grammatiktests bislang keine zufrieden stellenden Aussagen getroffen wurden. Informationen zur Konstruktvalidität des DSH-Grammatiktests erhebe ich in einer Studie, die in Kapitel 4.2 (Seite 114 ff) vorgestellt wird.

Authentizität und Interaktivität: Überlegungen zur Authentizität und Interaktivität von Grammatiktests sind nicht mit dieser Terminologie angestellt worden. Mit einer anderen Terminologie wurden jedoch vergleichbare Konzepte diskutiert. Rea-Dickins richtet – ganz unter dem Eindruck der kommunikativen Methode – an Grammatiktests die Frage: "What makes a grammar test communicative?" (Rea-Dickins, 1991: 112). Kieweg (siehe unten) unterscheidet zwischen passivem, produktivem und anwendungsorientiertem Grammatikwissen. Ich stelle beide Ansätze vor und beziehe sie auf den DSH-Grammatiktest.

Rea-Dickins stellt folgende Anforderungen an einen kommunikativen Grammatiktest:

> [...] five factors that contribute to the 'communicative' nature of a grammar test. These include:

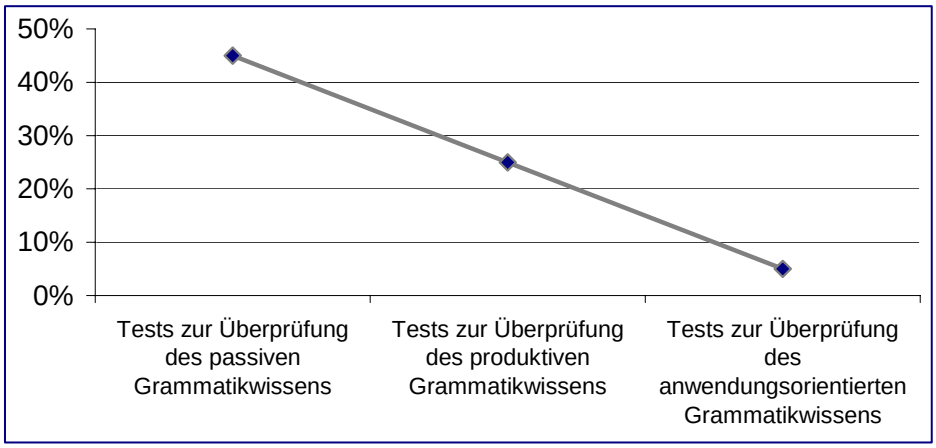
1. the contextualisation of test items: a test should not comprise a number of decontextualised single sentences;
2. the identification of a communicative purpose for the test activity;
3. the identification of an audience to whom the communication is addressed;
4. instructions to the test taker that focus on meaning rather than on form;
5. the opportunity for the test taker to create his/her own message and to produce grammatical responses as appropriate to a given context (Rea-Dickins, 1991: 125).

Derartige Kriterienkataloge, die in vergleichbarer Form auch von anderen aufgestellt wurden (z. B. Morrow, 1979), sind nicht unumstritten; auch der Terminus "kommunikativer Grammatiktest" hat keine Verbreitung gefunden (Alderson, 1981: 48; Davies, 1991; Rea-Dickins, 2001). Dennoch ist es sinnvoll zu fragen, ob der DSH-Grammatiktest im Sinne dieses Kriterienkatalogs als "kommunikativer" Grammatiktest anzusehen ist. Das einzige Merkmal, das vom DSH-Grammatiktest erfüllt wird, ist die Kontextualisierung der Items. Andere Merkmale werden nicht erfüllt: Warum soll der Text verändert werden? Soll er besonders verständlich werden? Soll er in eine wissenschaftliche Sprache gesetzt werden? Nur wenn derartige Aspekte eine Rolle spielten, könnte man von einem "kommunikativen" Grammatiktest sprechen. In ähnlicher Weise müsste man die Frage beantworten, ob der DSH-Gramatiktest über "authentische" oder "interaktive" Züge verfügt. Diese Konzepte sind verwandt (Übersicht in: Bachman, 2000: 3-4).

Ebenso wie beim Unterricht lassen sich bei Tests kommunikative Ansätze von formalen oder funktionalen Grammatikansätzen abgrenzen (Rea-Dickins, 1991: 113-114). In einem Unterricht mit einem formalen Grammatikansatz (*focus on formS*) fragt man: Was bedeutet diese Form? Ein funktionaler Grammatikansatz geht von der Frage aus: Wie kann diese Absicht ausgedrückt werden? Ein kommunikatives Grammatikverständnis befasst sich neben diesen beiden Ebenen noch mit der pragmatischen Angemessenheit und den Auswirkungen auf Form und Funktion.

Diese Unterscheidungen ähneln Kategorien, die Kieweg (1999) verwendet: Er unterscheidet zwischen "passivem", "produktivem" und "anwendungsorientiertem" Grammatikwissen. Ob die passive, produktive oder anwendungsorientierte Beherrschung der Grammatik geprüft wird, hat Auswirkungen auf den Schwierigkeitsgrad und auf die Leistungen. Dies geht aus Untersuchungen hervor, auf die Kieweg verweist. Er machte das Thema "present perfect/present perfect progressive" zum Gegenstand verschiedener Tests (Kieweg, 1999: 4-6). Das passive Grammatikwissen wurde mit Hilfe von Multiple-Choice Aufgaben geprüft, bei denen der sprachlich korrekte Ausdruck ausgewählt werden sollte. Das produktive Grammatikwissen sollte mit einem Lückentext

geprüft werden, bei dem die Zeitform mit einem vorgegebenen Verb gebildet werden musste. Die Aufgabe zum anwendungsorientierten bzw. interaktiven Grammatikwissen bestand aus einem unvollständigen Dialog, den die Schüler situationsadäquat vervollständigen sollten. (Eine Aufgabe, welche wohl nicht allen Anforderungen entspricht, die Rea-Dickins an einen kommunikativen Grammatiktest stellt, ihnen aber nahe kommt.) Die Testergebnisse verschlechtern sich in Bezug auf das Grammatikthema "present perfect/present perfect progressive", wenn nicht nur das Erkennen der richtigen Struktur, sondern auch der Gebrauch getestet wird (siehe Abbildung 9, Seite 89). Der Test des anwendungsorientierten Grammatikwissens habe "in vielen Klassen beinahe bis zum vollständigen Zusammenbruch" geführt (Kieweg, 1999: 5).



nach: Kieweg 1999: 5.

Abbildung 9: Leistungen in Tests von passivem, produktivem und interaktivem Grammatikwissen

Im "DSH-Handbuch für Prüferinnen und Prüfer" wird darauf hingewiesen, dass in DSH-Grammatiktests sowohl rezeptiv als auch produktiv mit wissenschaftssprachlich relevanten Strukturen umgegangen wird, d. h. passives und produktives Grammatikwissen geprüft wird (FaDaF, 2001: 7/2). Die Struktur, mit der eine Transformation vorgenommen werden soll, muss zunächst erkannt werden (passives Grammatikwissen). Bewertet wird das produktive Grammatikwissen. Obwohl die Transformationsaufgaben des DSH-Grammatiktests durch die konsequente Kontextualisierung komplexer sind als isolierte Aufgaben zum produktiven Grammatikwissen, ist es nicht angebracht, von einem anwendungsorientierten Test des Grammatikwissens zu sprechen (ebenso wie man nicht von einem kommunikativen Grammatiktest sprechen kann), sondern allein von einem Test des produktiven Grammatikwissens. Auf der Skala "Test des passiven – produktiven – anwendungsorientierten Grammatikwissen" dürfte der DSH-Grammatiktest jedoch näher am "anwendungsorientierten Wissen" liegen als die jeweiligen Prüfungsteile des KDS und der ZOP, bei denen die Kandidaten explizite Informationen über die zu verwendende Struktur erhalten.

Ebenso wie man beim DSH-Grammatiktest kaum von einem kommunikativen Test oder von einem Test des anwendungsorientierten Grammatikwissens sprechen kann, so sind auch die Authentizität und die Interaktivität des DSH-Grammatiktests als gering einzuschätzen.

Auswirkungen: Grundsätzliche Überlegungen zu den Auswirkungen von Sprachtests wurden in Kapitel 2.2 angestellt (Seite 56 f). Hier möchte ich darüber nachdenken, welche Auswirkungen vom DSH-Grammatiktest zu erwarten sind und wie diese zu bewerten sind.

Wenn von Testauswirkungen die Rede ist, wird damit noch keine Aussage über die Qualität getroffen. Der Begriff ist wertfrei. Wann ist eine Testauswirkung als wünschenswert, positiv anzusehen, wann als negativ? Man könnte den Standpunkt vertreten, dass Aktivitäten zum Spracherwerb in jedem Fall positiv sind, unabhängig davon, ob sie durch einen bevorstehenden Test ausgelöst wurden oder nicht. Der Lerneinsatz sollte meiner Ansicht nach die Kandidaten allerdings auch dem Ziel näher bringen, die studienbezogenen Sprachkenntnisse zu verbessern. Negative Auswirkungen könnten sich beispielsweise ergeben, wenn Inhalt und Konstruktion des Tests nicht mit den Lehr- und Lernzielen bzw. mit der angestrebten Fertigkeit übereinstimmen. Dann unterscheidet

sich die Vorbereitung auf den Test von der Vorbereitung auf die angestrebte Sprachverwendungssituation. Die Lerner würden sich Fähigkeiten aneignen, mit denen sie zwar in der Testsituation bestehen, nicht aber in der realen Sprachverwendungssituation. Wenn die Vorbereitung auf den Test und die Vorbereitung auf die angestrebte Sprachverwendungssituation hingegen identisch sind, kann man von einer positiven Auswirkung ausgehen. Demnach ist bei einem direkten Test eher von einer positiven Auswirkung auszugehen als bei einem indirekten Test.

Nach Messick (1996) führen derartige Überlegungen jedoch leicht in die Irre: Seiner Ansicht nach entstehen negative Testauswirkungen vor allem, wenn die Konstruktvalidität gering ist. Wenn der Test wichtige Aspekte des Testkonstrukts nicht berücksichtigt und die Testteilnehmer relevante Komponenten ihrer sprachlichen Leistungsfähigkeit nicht unter Beweis stellen konnten, erzielen sie möglicherweise ein Ergebnis, das ihrem wahren Leistungsstand nicht entspricht. Dies kann zu zwei Fehleinschätzungen mit negativen Auswirkungen führen: Kandidaten erzielen ein zu niedriges Ergebnis und müssen die negativen Folgen hinnehmen oder sie erzielen ein zu hohes Ergebnis und geraten in Situationen, denen sie sprachlich nicht gewachsen sind. Messick fasst zusammen:

> It would seem, then, that if one is concerned with fostering positive washback and reducing negative washback, one should concentrate first on minimizing construct under-representation and construct-irrelevant difficulty in the assessment. That is, *rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback* (1996: 252; Hervorhebungen im Original).

Diese Argumentation verfolgt Messick auch als Mitarbeiter des *Educational Testing Service* (ETS), dem Testinstitut, das für den TOEFL zuständig ist. Beim TOEFL legt man wie auch beim TestDaF relativ wenig Wert auf die Authentizität der Texte, auf die Offenheit der Fragestellungen oder auf die Verwirklichung von direkten Testverfahren.

Wenn man die Nützlichkeit eines Tests einschätzen möchte, sind Testauswirkungen meiner Ansicht nach jedoch als eigenständige Komponente auch unabhängig von der Konstruktvalidität zu berücksichtigen. Beim DSH-Grammatiktest muss etwa überlegt werden, wie die Kandidaten zur Beherrschung der Grammatik geführt werden, die sie im Test unter Beweis stellen sollen. Im Abschnitt über explizites und implizites Grammatikwissen (Seite 76) wurde bereits argumentiert, dass durch en DSH-Grammatiktest nicht festgelegt ist, welcher Lernweg beschritten, welche Unterrichtsmethode gewählt werden soll. Denkbar ist ein Unterricht, der eine explizite Grammatikvermittlung ver-

meidet. Hier wird die Diskussion um den Grad der Explizitheit der Sprachvermittlung berührt (Diehl *et al*., 2000; Doughty/Williams, 1998; Eckerth, 2000; Green/Hecht, 1992; Hinkel/Fotos, 2001; Purpura, 2004; Rall, 2001). In einem derartigen Unterricht würden die Lerner ihr Grammatikwissen auf natürlichem Wege erweitern und ihre Lernersprache mit diesem Wissen erweitern (zum *Natural Approach* siehe Krashen, 1982; Krashen/Terrell, 1983; Prabhu, 1987). Möglich wäre auch das andere Extrem: Unterricht nach der Grammatik-Übersetzungsmethode (*focus on formS*). Möglich ist ebenfalls eine Grammatikvermittlung im Rahmen eines primär inhaltlich orientierten Unterrichts, wie bei *focus on form*.

Diese Unterscheidungen möchte ich kurz erläutern: In der englischsprachigen Sprachwissenschaft hat man das Begriffspaar "*focus on form*" und "*focus on formS*" geprägt. Während man unter "*focus on formS*" einen Unterricht versteht, der grammatische Phänomene isoliert von inhaltlichen Zusammenhängen thematisiert, soll das inhaltliche Thema bei der Reflexion über sprachliche Strukturen bei einem Unterricht nach dem Prinzip "*focus on form*" nicht aus dem Blick verloren werden (del Pilar García Mayo, 2002; Doughty/Williams, 1998; Ellis/Basturkmen/Loewen, 2002; Hyland, 2003). Die folgenden Definitionen verdeutlichen theoretische und praktische Ansätze dieses zweiten Prinzips:

> ...focus on *form* ... overtly draws students' attention to linguistic elements as they arise incidentally in lessons whose overriding focus is on meaning or communication (Long, 1991: 45-46; zit. n. Doughty/Williams, 1998: 3; Hervorhebung im Original).
>
> ...focus on form often consists of an occasional shift of attention to linguistic code features – by the teacher and/or one or more students – triggered by perceived problems with comprehension or production (Long/Robinson, 1998: 23).

Deutlich wird: "*Focus on form*" bezieht sich nicht auf die Unterscheidung zwischen expliziter und impliziter Grammatikvermittlung, sondern beschreibt ein Unterrichtsmodell, bei dem situationsabhängig die eine oder andere Vorgehensweise beschritten wird. Dieser Ansatz ist meiner Ansicht nach durch die Erkenntnisse der Spracherwerbsforschung und der empirischen Unterrichtsforschung abgesichert: Lerner erzielen einen Lernzuwachs, wenn sie sich mit sprachlichen Phänomenen aus der Grammatik beschäftigen, welche ihrer Lernstufe angemessen sind. Eine Vermittlung, die auch explizit vorgeht, hat sich als effektiver erwiesen. Eine Konzentration auf die Grammatikvermittlung ohne Einbindung in inhaltliche Zusammenhänge und damit auch ohne eine Erweiterung des Wortschatzes und ohne Erweiterung kommunikativer Fähigkeiten ist dagegen weniger produktiv (zur Diskussion um Grammatikvermittlung siehe Butzkamm, 1995;

Doughty/Williams, 1998; Harden/Marsh, 1993; Hinkel/Fotos, 2001; Norris/Ortega, 2000; Rall, 2001; Zimmermann, 1990; Zimmermann/Wißner-Kurzawa, 1985). Ich teile diese Einschätzung und würde von *negativen* Auswirkungen eines Grammatiktests ausgehen, wenn nachzuweisen wäre, dass er zu einer isolierten Konzentration auf die Grammatik (im Sinne von "*focus on formS*") führt.

Ich komme zurück zum DSH-Grammatiktest: Welche Testvorbereitung legt der DSH-Grammatiktest nahe? Mit der Existenz eines Grammatiktests wird ein Signal gesetzt. Es lautet: Grammatik stellt einen Teil der sprachlichen Vorbereitung auf ein Studium dar. Durch den Verzicht auf Metasprache und durch die Kontextualisierung wird ein Unterricht ermöglicht, in dem Grammatik im Sinne von "*focus on form*" in Verbindung mit Inhalten und mit Aussageabsichten vermittelt wird. (Etwas anders dürfte das Signal im Falle des Grammatiktests im KDS oder im ZOP interpretiert werden, da die Phänomene geordnet dargeboten werden und da Metasprache verwendet wird: Die Testteilnehmer müssen auch die Terminologie verstehen. Daher müssen Termini der Sprachwissenschaft im prüfungsvorbereitenden Unterricht vermittelt werden.) Empirische Studien zu den Auswirkungen von Grammatiktests sind mir nicht bekannt. Eigene Studien zu diesem Thema werden in Kapitel 4.4 (Seite 158 ff) beschrieben.

Ökonomie: Die Ökonomie eines Sprachtests wurde in Kapitel 2.2 (Seite 59 f) nach Bachman und Palmer definiert als das "Verhältnis zwischen den Mitteln, die für die Konzeption, Entwicklung und den Einsatz des Tests benötigt werden und den Mitteln, die dafür zur Verfügung stehen". Es wurde bereits darauf hingewiesen, dass der DSH-Grammatiktest ein zusätzlicher Prüfungsteil ist, der die Prüfung verlängert und sie damit weniger ökonomisch macht. Mit der Überarbeitung der DSH-Rahmenordnung 2004 wurde die Rolle allerdings eingeschränkt: Der DSH-Grammatiktest wurde gekürzt und an das Leseverstehen angegliedert. Er wird jedoch nach wie vor gesondert bewertet.

Zusammenfassung und Ausblick

In diesem Kapitel wurde der DSH-Grammatiktest anhand von sprachwissenschaftlichen und testtheoretischen Konzepten untersucht. Die Einschätzungen wurden mit Verweis auf Forschungsergebnisse, argumentativ oder auch spekulativ begründet. Dabei wurden offene Fragen identifiziert.

Der DSH-Grammatiktest ist ein indirekter Kompetenztest, der auch explizites Grammatikwissen hervorruft. Sprachentwicklungsstufen werden nicht berücksichtigt. Den typischen Schwierigkeiten indirekter Kompetenztests (Testkriterium und Interpretation der Ergebnisse unklar, wenig Hinweise auf die Festsetzung des Schwellenwertes) dürfte in der Praxis begegnet werden, indem er wie eine Kursabschlussprüfung eingesetzt wird. Dies ist jedoch eine unangemessene Vorgehensweise für einen Sprachtests für den Hochschulzugang.

Die Reliabilität des DSH-Grammatiktests ist nicht gesichert. Dies bezieht sich vor allem auf die Paralleltestreliabilität. Unklar ist, welche Rolle unterschiedliche Testmethoden-Merkmale auf die Reliabilität haben. Während die theoriegeleitete Bestimmung des Testkonstrukts möglich ist ("produktive Grammatikkompetenz"), ist die Frage, welche Interpretationen darüber hinaus noch möglich und sinnvoll sind, empirisch nicht hinreichend belegt.

Der DSH-Grammatiktest ist weder als besonders authentisch noch als besonders interaktiv anzusehen. Über Testauswirkungen liegen keine Untersuchungen vor. Obwohl der DSH-Grammatiktest relativ einfach zu erstellen und zu korrigieren ist, führt die Existenz der Items zur Grammatik zu einer Verlängerung der Prüfung. Dies ist nur zu rechtfertigen, wenn der DSH-Grammatiktest zur Nützlichkeit der DSH beitragen würde. Dies ist jedoch, so lautet mein Fazit am Ende dieses Kapitels, bislang nicht deutlich geworden.

Aus den Überlegungen in diesem Kapitel leite ich folgende Fragen ab, die sich bislang noch nicht zufrieden stellend beantworten lassen: Welche Auswirkungen haben unterschiedliche Testmethoden-Merkmale von Grammatiktests auf die Paralleltestreliabilität? Das Testkonstrukt ist "produktive Grammatikkompetenz". Welche weiteren Interpretationen sind möglich? Welche Auswirkungen sind zu erwarten? Die Nützlichkeit und die Legitimität des DSH-Grammatiktests hängen meiner Ansicht nach von diesen zentralen Fragen ab, welche ich im folgenden Kapitel aufgreife.

4. Studien zum DSH-Grammatiktest

Übersicht: Kapitel 4
Im Folgenden stelle ich Studien zu Grammatiktests vor, die sich auf vier Fragenkomplexe beziehen: Die Auswirkungen unterschiedlicher Testmethoden-Merkmale (Paralleltestreliabilität), die Konstruktvalidität von Grammatiktests, die Auswirkungen des DSH-Grammatiktests auf die Zulassungsentscheidung und schließlich die Auswirkungen des Grammatiktests auf Lehr- und Lernprozesse. Im Hintergrund steht die Frage nach der Legitimität des DSH-Grammatiktests. In Kapitel 4.5 fasse ich die Untersuchungen zusammen und ziehe Schlussfolgerungen für den Umgang mit Grammatik in Sprachtests für den Hochschulzugang.

Die Studien zu Grammatiktests ergeben sich aus den Überlegungen in Kapitel 3. Die Studien greifen offene Fragen zur Reliabilität, zur Konstruktvalidität und zu den Auswirkungen des DSH-Grammatiktests auf. Ausgangspunkt für die Studien ist der Trend, auf Prüfungsteile zur Grammatik in Sprachtests für den Hochschulzugang zu verzichten. Beim IELTS wurde ein Grammatiktest, der in einer Pilotversion enthalten war, nach einer Überarbeitung gestrichen. Beim TestDaF kommt man ebenfalls mit vier Prüfungsteilen zu den sprachlichen Fertigkeiten Hören, Lesen, Schreiben und Sprechen aus. Beim TOEFL wird man in Zukunft auf den Prüfungsteil "Structure and Written Expression" verzichten und ebenfalls die vier sprachlichen Fertigkeiten in den Prüfungsteilen abbilden.

Auch bei der DSH wurde bei der Überarbeitung in den Jahren 2003 und 2004 das Format der Prüfung geändert. Der Umfang des DSH-Grammatiktests wurde verringert, die

Gewichtung reduziert. Der DSH-Grammatiktest soll an den Prüfungsteil Leseverstehen angegliedert werden. Diese Maßnahmen zeigen ein gewisses Unbehagen mit dem Grammatiktest: Er wird im Prüfungsdickicht versteckt. Sollte man ihn ganz streichen? Ist er obsolet? Die DSH würde durch einen Verzicht auf den Grammatiktest ökonomischer. Sie würde kaum an Interaktivität oder an Authentizität verlieren. Der DSH-Grammatiktest dürfte dazu beitragen, dass die DSH häufig wie ein Kursabschlusstest eingesetzt wird. Ein Verzicht darauf würde den Charakter der Feststellungsprüfung betonen, was ich für sinnvoll hielte.

Die Erfahrungen mit IELTS und TOEFL können nicht direkt auf die DSH übertragen werden. Die DSH findet unter anderen Voraussetzungen statt, so dass unterschiedliche Vorgehensweisen zulässig oder sogar notwendig sind. Die Annahmen, die zum Verzicht auf den expliziten Grammatiktest geführt haben, und der Nutzen, der damit verbunden ist, treffen nicht in gleichem Maße auf die DSH zu. Erstens ist die Übertragung der Untersuchungsergebnisse der IELTS-Studie (Alderson, 1993) auf die DSH fragwürdig. Der Grammatiktest der Pilotversion von IELTS gleicht dem Testteil "Wissenschaftssprachliche Strukturen" der DSH zwar durch die Vorgabe der Kontextualität. Allerdings liegt der Schwerpunkt bei der DSH auf grammatischen Strukturen und betrifft weniger die lexikalischen Probleme, so dass die Untersuchungsergebnisse nicht direkt übertragen werden können. Zweitens trifft die Interpretation der Ergebnisse im Fall des DSH-Grammatiktests nicht auf die typischen Schwierigkeiten wie bei anderen indirekten Kompetenztests. Die DSH wird dezentral von Sprachlehrerinnen und –lehrern vorgenommen, welche die Studiensituation in den verschiedenen Fachbereichen überblicken und viele Testteilnehmer aus dem Unterricht kennen (diese Argumentation vertritt z. B. Wintermann, 1998: 110). Möglicherweise können sie aufgrund ihrer Erfahrungen interpretieren, mit welchem Ergebnis ein Prüfungsteil als bestanden gilt. Drittens dürfte das Argument der Testökonomie im Fall der DSH nicht in gleichem Maße zum Tragen kommen. Häufig werden Grammatiktests gerade wegen der ökonomischen Vertretbarkeit in Sprachtests aufgenommen, sie lassen sich z. B. problemlos auswerten. Der TestDaF wird von einem Team zentral gestellt und ausgewertet, in die Erstellung kann mehr Sorgfalt einfließen, als bei dezentralen Prüfungen; beim TestDaF kann beispielsweise mittels aufwändiger Probedurchläufe eine Standardisierung sichergestellt werden. Bei der dezentralen DSH können allenfalls kleine Kontrollgruppen gefunden werden, Erprobungen des Tests stoßen auf organisatorische Probleme. Unter

diesen Umständen wäre ein Verzicht auf einen Testteil, der einfach zu erstellen und zu bewerten ist, schon allein mit dem Hinweis auf die Testökonomie fragwürdig.

Es gibt aber offene Fragen, die bislang nicht argumentativ beantwortet werden konnten: Die geringe Reliabilität ist die Schwäche der DSH. Erstens muss also gefragt werden, welchen Beitrag der DSH-Grammatiktest zur Reliabilität leistet (Kapitel 4.1). Es stellt sich zweitens die Frage nach dem Wert des DSH-Grammatiktests zum Testkonstrukt eines Sprachtests für den Hochschulzugang (Kapitel 4.2). Drittens sind die Auswirkungen des DSH-Grammatiktests zu untersuchen. Hier interessiert mich, welche Auswirkungen der DSH-Grammatiktest auf die Zulassungsentscheidung hat (Kapitel 4.3) und welche Auswirkungen auf die Testvorbereitung zu erwarten sind (Kapitel 4.4). Mit den Studien soll nicht allein die Frage nach einem möglichen Verzicht auf den DSH-Grammatiktest im Sinne eines Alles-oder-Nichts beantwortet werden. Die Studien sollen vor allem zu mehr Klarheit im Umgang mit Grammatiktests führen.

4.1. Unterschiedliche Testmethoden-Merkmale

Übersicht: Kapitel 4.1
Welche Auswirkungen unterschiedliche Testmethoden-Merkmale haben, wird anhand von drei unterschiedlichen DSH-Grammatiktests erhoben: Der erste Grammatiktest ist dadurch gekennzeichnet, dass die sprachlichen Phänomene geordnet behandelt und mit Beispiel und linguistischer Metasprache erläutert werden; im zweiten Test wird auf eine Ordnung verzichtet, nicht aber auf metasprachliche Erläuterungen; im dritten (dem DSH-Prototyp) werden Transformationsaufgaben ohne Ordnung und Metasprache gestellt.

4.1.1. Fragestellung und Methode

Wenn Grammatiktests über gleiche Testmethoden-Merkmale verfügen, steht die Reliabilität meiner Ansicht nach nicht in Frage. Bei Grammatiktests dürfte die Itemschwierigkeit beispielsweise weniger vom Thema abhängen als bei Tests zum Leseverstehen oder zum Hörverstehen.

Bei DSH-Grammatiktests ist dies jedoch nicht gewährleistet: Dass sich Grammatiktests in Sprachtests für den Hochschulzugang in durchaus wesentlichen Aspekten unterscheiden, ging aus der Beschreibung in Kapitel 3 hervor. *Den* DSH-Grammatiktest gibt es nicht. Der Prototyp aus dem DSH-Handbuch dürfte zwar häufig als Muster genommen werden, es ist aber davon auszugehen, dass verschiedene Formate eingesetzt werden. Daher sollte die Analyse unterschiedlicher Testmethoden-Merkmale am Beginn von Studien zu Grammatiktests stehen.

Zwei Besonderheiten lassen sich beobachten: der Einsatz von Metasprache (Beispiel: "Verwenden Sie in Ihrer Antwort das Passiv!") und die Ordnung der Aufgaben nach sprachlichen Phänomenen (Beispiel: "Ändern Sie die Sätze nach folgendem Muster...").

Beide Merkmale sind gemäß DSH-Rahmenordnung möglich, laut DSH-Handbuch jedoch nicht vorgesehen. Musterprüfungen von DSH-Ausrichtern ist jedoch zu entnehmen, dass Prüfungen mit diesen Anweisungen durchaus eingesetzt werden. Die Eigenschaften "Einsatz von Metasprache" und "Ordnung der Aufgaben" werden durchgängig in anderen Sprachtests wie z. B. dem KDS oder der ZOP verwendet (siehe Kapitel 3.1, Seite 70 f).

Es daher von Interesse, ob der Einsatz von Metasprache und die Ordnung der Aufgaben einen Einfluss auf die Ergebnisse haben und die Konstruktvalidität beeinflussen. Indikatoren wären unterschiedliche mittlere Schwierigkeitsgrade, Varianzen oder Kovarianzen. Die Hypothese lautet, dass unterschiedliche Formate durchaus zu unterschiedlichen Ergebnissen führen, dass also ein Methodeneffekt zu beobachten ist. Sollte sich ein starker Methodeneffekt bestätigen, müssten die Vorgaben für die Erstellung von Grammatiktests noch enger gefasst werden.

Fragestellung
Welche Auswirkungen haben unterschiedliche Testmethoden-Merkmale von Grammatiktests (mit/ohne Metasprache, mit/ohne Ordnung der sprachlichen Phänomene)?

Es handelt sich bei den Studien zu den Grammatiktests in Kapitel 4 um zusammenhängende Untersuchungen, die jeweils mit Blick auf eine Fragestellung interpretiert werden. Um den Methodeneffekt zu erfassen, setzte ich drei unterschiedliche DSH-Grammatiktests ein, welche von Teilnehmern des Studienkollegs an der Fachhochschule Konstanz und von Teilnehmern eines DSH-Vorbereitungskurses bearbeitet wurden. Die Tests wurden im Rahmen des Deutschunterrichts durchgeführt. Die Testteilnehmerinnen und -teilnehmer hatten für die Bearbeitung jeweils ca. 45 Minuten Zeit. Die Tests wurden von mir korrigiert und anschließend im Unterricht besprochen. In einem Zeitraum von maximal sechs Wochen bearbeiteten 110 Kollegiaten[2] die Grammatiktests, wobei nicht alle an den drei Tests teilnahmen. Die Reihenfolge, in der die

[2] Bei "Kollegiaten" handelt es sich um Teilnehmerinnen und Teilnehmer eines Ausländerstudienkollegs.

Tests bearbeitet wurden, war unterschiedlich. Trotz der zeitlichen Nähe ist ein Lerneffekt nicht auszuschließen.

Es wurden folgende Grammatiktests ausgewählt (siehe auch Tabelle 6, Seite 101):

- Der Grammatiktest "Flurbereinigung" ist ein Mustertest aus dem DSH-Handbuch, daher bezeichne ich ihn auch als "Prototyp" (siehe Abbildung 7, Seite 74). Im Test wird keine Metasprache verwendet, die sprachlichen Phänomene werden ungeordnet behandelt und müssen von den Testteilnehmern erkannt werden.

- Der Grammatiktest "Teilzeitarbeit" wurde als Mustertest von einem DSH-Ausrichter im Internet veröffentlicht. Der Test verwendet Metasprache (siehe Abbildung 10, Seite 103).

- Auch der Grammatiktest "Neue Medien" wurde als Mustertest von einem DSH-Ausrichter im Internet veröffentlicht. Der Test verwendet Metasprache. Die sprachlichen Phänomene werden außerdem geordnet angeboten (siehe Abbildung 11, Seite 103).

Die drei Tests verlangen sprachliche Transformationsaufgaben. Sie erfüllen die Anforderungen aus der neuen DSH-Rahmenordnung, allerdings sind sie nicht mit dem Leseverstehen verschmolzen, sondern – wie in der alten DSH-Rahmenordnung vorgesehen – als eigenständiger Prüfungsteil konzipiert (HRK/KMK, 2004). Die Vergleichstests "Neue Medien" und "Teilzeitarbeit" orientieren sich nicht an der Musterprüfung und den Anweisungen aus dem DSH-Handbuch. Wenn man die von DSH-Ausrichtern im Internet veröffentlichten Musterprüfungen als repräsentativ für die tatsächlich durchgeführten Prüfungen ansieht, dann sind mit dieser Auswahl viele DSH-Grammatiktests erfasst. Nur wenige Ausrichter setzen völlig andere Tests unter der Überschrift "Wissenschaftssprachliche Strukturen" ein, wie z. B. C-Tests oder Aufgaben zur Lexik ("Erklären Sie folgende Begriffe aus dem Text ..."). Dies scheinen Ausnahmen zu sein. Die für die Studie ausgewählten Tests werden im Folgenden genauer beschrieben. Eine Übersicht bietet Tabelle 6 (Seite 101).

Der Prototyp Grammatiktest "Flurbereinigung"

Der Grammatiktest "Flurbereinigung" stammt aus dem DSH-Handbuch. Der auf einem Text zur "Flurbereinigung" beruhende Test wurde in Kapitel 3 bereits beschrieben und eingeordnet (siehe Abbildung 7, Seite 74). Wenn man sich die im Internet veröffentlich-

ten Musterprüfungen der DSH-Ausrichter ansieht, stellt man fest, dass sich das Format des Grammatiktests aus dem Handbuch mehr und mehr durchzusetzen scheint. Erfragt wird ein relativ breites Spektrum an sprachlichen Strukturen, von denen jede jeweils nur ein- bis zweimal thematisiert wird (siehe Tabelle 6, Seite 101). Metasprache wird nicht verwendet, die sprachlichen Phänomene sind nicht geordnet.

Tabelle 6: Grammatiktests der Studie "Unterschiedliche Testmethoden-Merkmale"

	Prototyp Grammatiktest "Flurbereinigung"	**Grammatiktest mit Metasprache "Teilzeitarbeit"**	**Grammatiktest mit Ordnung "Neue Medien"**
Thema	Flurbereinigung	Teilzeitarbeit	Neue Medien
Ausgangstext	beibehalten	beibehalten	nur thematische Anlehnung an einen Ausgangstext
Anordnung der sprachlichen Phänomene	zufällig	zufällig	nach Phänomenen geordnet
Aufgabenstellungen	ohne metasprachliche Anweisungen	mit metasprachlichen Anweisungen	mit metasprachlichen Anweisungen
Beispiellösungen	ohne Beispiellösungen	ohne Beispiellösungen	mit Beispiellösungen
maximale Punktzahl	26,5	27	25
Items: *Anzahl der Items (Transformationsaufgaben)*	10 Sätze mit 11 Items mit mehrstufiger Bewertung	6 Sätze mit 9 Items mit mehrstufiger Bewertung	12 mit mehrstufiger Bewertung
Art der Items:			
• *Relativsatz → Partizip als Attribut*	1		2
• *Partizip als Attribut → Relativsatz*	2	1	1
• *Verbalstil → Nominalstil*	1	3	
• *Nominalstil → Verbalstil*	2		
• *Passiversatz → Passiv/unpersönliches "man"*	2	1	
• *Passiv → Passiversatz*		1	
• *Präpositionalphase → Nebensatz*	1	1	
• *Nebensatz → Präpositionalphrase*		1	3
• *Semantische Veränderung*	1		
• *Modalähnliches Verb → Modalverb*	1		3
• *Hypothese /Konjunktiv*			1
• *Satzgefüge*			2

Grammatiktests mit Metasprache "Teilzeitarbeit"

Der Vergleichstest mit Metasprache "Teilzeitarbeit" stammt aus einer Musterprüfung. Er basiert ebenfalls auf kontextualisierten Transformationsaufgaben, zusätzlich werden metasprachliche Anweisungen verwendet (siehe Tabelle 6, Seite 101 und Abbildung 10, Seite 103). Es werden nicht nur der Ausgangssatz und ein veränderter Satz mit Lücken präsentiert, sondern auch Hinweise zur Bearbeitung gegeben, wie "Nominalisierung", "Relativsatz", "Passiv", "Verbalisierung/Nebensatz", "Nebensatz". Die Kandidaten müssen nicht mehr selbst herausfinden, wie ein Satz geändert werden soll, es wird ihnen vielmehr in der Anweisung ausdrücklich mitgeteilt. Der Grammatiktest "Metasprache" verzichtet dabei – anders als der folgende Grammatiktest "Ordnung" – auf ein Beispiel zur Illustration der erwarteten Antwort. Von den Kandidaten wird verlangt, dass sie die Metasprache verstehen. Bisweilen lassen sich auch andere sinnvolle Transformationen bilden, ohne dass die Anweisung befolgt würde. In dem Fall können die Kandidaten nicht die volle Punktzahl erhalten. Der authentische Grammatiktest "Metasprache" unterscheidet sich damit in einem zentralen Merkmal von dem Grammatiktest "Flurbereinigung". Das Spektrum der thematisierten Strukturen ist etwas enger.

Grammatiktests mit Ordnung "Neue Medien"

Ein Format, das ebenfalls gelegentlich anzutreffen ist, ist die Ordnung der sprachlichen Phänomene (siehe Tabelle 6, Seite 101 und Abbildung 11, Seite 103). Die Kandidaten erhalten neben der Anweisung, welche in diesem Fall auch Metasprache verwendet, zusätzlich ein Beispiel, wie die folgenden Sätze zu verändern sind. Durch die Gruppierung der Items nach sprachlichen Gesichtspunkten muss ein textueller Zusammenhang in der Regel aufgelöst werden. Im Grammatiktest "Ordnung", der für die Studie ausgewählt wurde, ist nur ein loser thematischer Zusammenhang zu erkennen, nicht aber ein zugrunde liegender Text. Der Test enthält zu jeder Struktur drei Items. Das wäre nicht konform mit dem DSH-Handbuch, widerspricht der Rahmenordnung aber nicht unbedingt. Dadurch ist das Spektrum jedenfalls kleiner.

Vervollständigen Sie den Text auf den folgenden Seiten durch eine korrekte und sinnentsprechende Umwandlung der unterstrichenen Satzteile.

Wandel der Arbeitszeiten: Teilzeitarbeit

Teilzeitbeschäftigung ist noch überwiegend ein "weibliches" Phänomen: Die Gruppe der Teilzeitbeschäftigten besteht zu 89 % aus Frauen. Nur drei Prozent der erwerbstätigen Männer in Deutschland gehen einer Teilzeitbeschäftigung nach, und entsprechend sind nur rund 11 % aller Teilzeitbeschäftigten Männer.

1. Die mittlerweile von vielen deutschen Firmen angebotenen Teilzeitmodelle werden von Männern nur sehr schleppend angenommen. **(Relativsatz)**
Die Teilzeitmodelle, _____
_____ , werden von Männern nur sehr schleppend angenommen.

2. Dies hängt u.a. damit zusammen, dass Teilzeitarbeit auch nur zu einem Teilzeiteinkommen und zu einer Teilzeitrente führt, und dies ist für die meisten Männer noch nicht akzeptabel. **(Passiv)**
Dies hängt u.a. damit zusammen, dass Teilzeitarbeit auch nur zu einem Teilzeiteinkommen und zu einer Teilzeitrente führt, und dies
_____
_____
[...]

(vollständiger Test: Anhang 2, Seite 356)

Abbildung 10: Grammatiktest mit Metasprache "Teilzeitarbeit"

1. Teil:
Verwandeln Sie jeweils den Relativsatz in ein Partizip (Partizip I oder Partizip II) oder ein Adjektiv mit der Endung "–bar".
Beispiel: *Die Zahl der Informationen, die in Datenbanken gespeichert sind, wächst explosionsartig.*
Lösung: *Die Zahl der in Datenbanken gespeicherten Information wächst explosionsartig.*
Einige Kritiker warnen vor den Folgen der Kommunikationsrevolution, die derzeit weltweit abläuft.
.....
Skeptiker befürchten eine Flut von Informationen, die nicht mehr kontrolliert werden kann.
.....
Der Mensch vergisst sehr schnell wieder einen großen Teil der Informationen, die er aufgenommen hat.
..... [...]

(vollständiger Test: Anhang 3, Seite 358)

Abbildung 11: Grammatiktest mit Metasprache und Ordnung "Neue Medien"

Um die Ergebnisse in den drei Grammatiktests besser in Beziehung setzen zu können, wurden sie linear in eine gemeinsame Skala transformiert (Prozentwerte). Die Studie zu den Auswirkungen unterschiedlicher Formate enthält deskriptive Statistiken (Mittelwerte, Standardabweichungen) sowie eine Überprüfung des Zusammenhangs zwischen den Ergebnissen der drei Tests mittels Korrelationen und Übereinstimmungskoeffizienten.

4.1.2. Ergebnisse und Diskussion

Ergebnisse: Die Mittelwerte und die Streuung der Ergebnisse in den drei Grammatiktests "Flurbereinigung", "Teilzeitarbeit" und "Neue Medien" unterscheiden sich[3]. Die Testteilnehmer erreichten im DSH-Grammatiktest "Neue Medien" im Mittel das höchste Ergebnis (65 Prozent; siehe Tabelle 7). Niedriger war das mittlere Ergebnis im Grammatiktest mit Metasprache "Teilzeitarbeit" (56 Prozent), am niedrigsten im Prototyp aus dem DSH-Handbuch "Flurbereinigung" (50 Prozent). Mit *t*-Tests zu abhängigen Stichproben kann gezeigt werden, dass die Unterschiede zwischen den Mittelwerten signifikant sind (siehe Tabelle 8). Der Grammatiktest ohne Metasprache und ohne Ordnung "Flurbereinigung" differenzierte am stärksten zwischen den Ergebnissen (s = 23,7 %).

Anhand von Korrelationskoeffizienten lässt sich ermitteln, wie groß der Zusammenhang zwischen den Ergebnissen ist. Vergleicht man verschiedene Versionen eines Tests, so kann man anhand von Korrelationskoeffizienten Informationen über die Paralleltestreliabilität gewinnen (siehe Kapitel 2.2, Seite 43). Alle drei Korrelationen zwischen den Ergebnissen in den Grammatiktests weisen auf mittlere und hohe Zusammenhänge (Tabelle 9). Sie sind signifikant. Am höchsten ist der Zusammenhang zwischen den Grammatiktests "Flurbereinigung" und "Neue Medien" ($r = 0{,}681$; $n = 135$; $p < 0{,}01$). Auf mittlere Zusammenhänge verweisen die Korrelationskoeffizienten nach Pearson zwischen den Grammatiktests "Flurbereinigung" und "Teilzeitarbeit" ($r = 0{,}437$; $n = 114$; $p < 0{,}01$) bzw. "Neue Medien" und " Teilzeitarbeit" ($r = 0{,}559$; $n = 128$; $p < 0{,}01$). Der höchste Zusammenhang besteht demnach zwischen den Grammatiktests "Ordnung – Neue Medien" und "Prototyp – Flurbereinigung", der niedrigste zwischen den Grammatiktests "Metasprache – Teilzeitarbeit" und "Prototyp – Flurbereinigung". Die Paralleltestreliabilität der Tests liegt damit zwischen 43 und 68 Prozent. Umgekehrt bedeutet dies: Sie sind zwischen 57 und 32 Prozent nicht reliabel.

Wie unterschiedlich die Tests sind, lässt sich weiter erfassen, indem man vergleicht, welche Ergebnisklassen die Kandidaten in den drei Tests erzielen. Nach der neuen

[3] Mit "Mittelwert" oder "mittlerem Ergebnis" wird in dieser Arbeit stets das arithmetische Mittel bezeichnet.

Rahmenordnung gibt es vier Ergebnisklassen: "unter DSH-1" oder "nicht bestanden", "DSH-1", "DSH-2" und "DSH-3". DSH-1 erzielt man mit einem Ergebnis ab 57 Prozent, DSH-2 ab 67 Prozent und DSH-3 ab 82 Prozent (HRK/KMK, 2004). Diese Ergebnisklassen beziehen sich auf das Gesamtergebnis, dennoch ist das Ergebnis im einzelnen Subtest relevant. Wenn das Gesamtergebnis der Schriftlichen und Mündlichen Prüfungen über 67 Prozent liegt, ist der Nachweis deutscher Sprachkenntnisse für die uneingeschränkte Zulassung erbracht. Von Interesse ist also vor allem, wie viele Kandidaten in den Grammatiktests mehr als 67 Prozent (also DSH-2 bzw. DSH-3) und wie viele weniger als 67 Prozent erzielten.

Welche Ergebnisklassen die Testteilnehmer in den drei Grammatiktests erzielten, geht aus der Abbildung 12 (Seite 107) hervor. Demnach erzielte ein Viertel der Kandidaten (24 %) im Grammatiktest "Flurbereinigung" das Ergebnis DSH-2 bzw. DSH-3, ein Drittel (33 %) erzielte DSH-2/DSH-3 im Grammatiktest mit Metasprache "Teilzeitarbeit" und etwa die Hälfte (49 %) erreichte DSH-2/DSH-3 im Grammatiktest mit Ordnung und Metasprache "Neue Medien".

Mit Übereinstimmungskoeffizienten lässt sich erfassen, wie viele Kandidaten in zwei (oder mehr) Tests vergleichbare Ergebnisse erzielen (siehe Abbildung 12, Seite 107). Die berechneten Übereinstimmungskoeffizienten umfassen den Anteil der Kandidaten, die in beiden Tests unter 67 Prozent erzielten und den Anteil der Kandidaten, deren Ergebnisse in beiden Tests über 67 Prozent lagen. Die Übereinstimmungskoeffizienten liegen zwischen 66 und 75 Prozent. Das bedeutet, dass zwei Drittel bis drei Viertel der Kandidaten die sprachliche Studierfähigkeit entweder in beiden Tests erreichten oder in beiden Tests nicht erreichten, dass sie also in beiden Tests mehr oder aber in beiden Tests weniger als 67 Prozent erzielten. Im Umkehrschluss bedeutet dies auch, dass ein Viertel bis ein Drittel der Kandidaten nach einem Test nicht studierfähig sind, nach dem anderen jedoch schon.

Tabelle 7: Grammatiktests – statistische Kennzahlen

DSH-Grammatiktests	**Prototyp "Flurbereinigung"**	**"Teilzeitarbeit" Metasprache**	**"Neue Medien" Ordnung**
Anzahl (N)	146	135	156
Mittelwert (*AM*) in %	50 %	56 %	65 %
Median (Wert, der die Gesamtzahl in zwei Hälften teilt; *Md*)	51 %	56 %	67 %
Standardabweichung (s) in %	23,7 %	19,2 %	20,3 %
Min. – Max. in %	0-100	12-96	13-98

Tabelle 8: Grammatiktests – Vergleich der Mittelwerte mit t-Tests zu abhängigen Stichproben

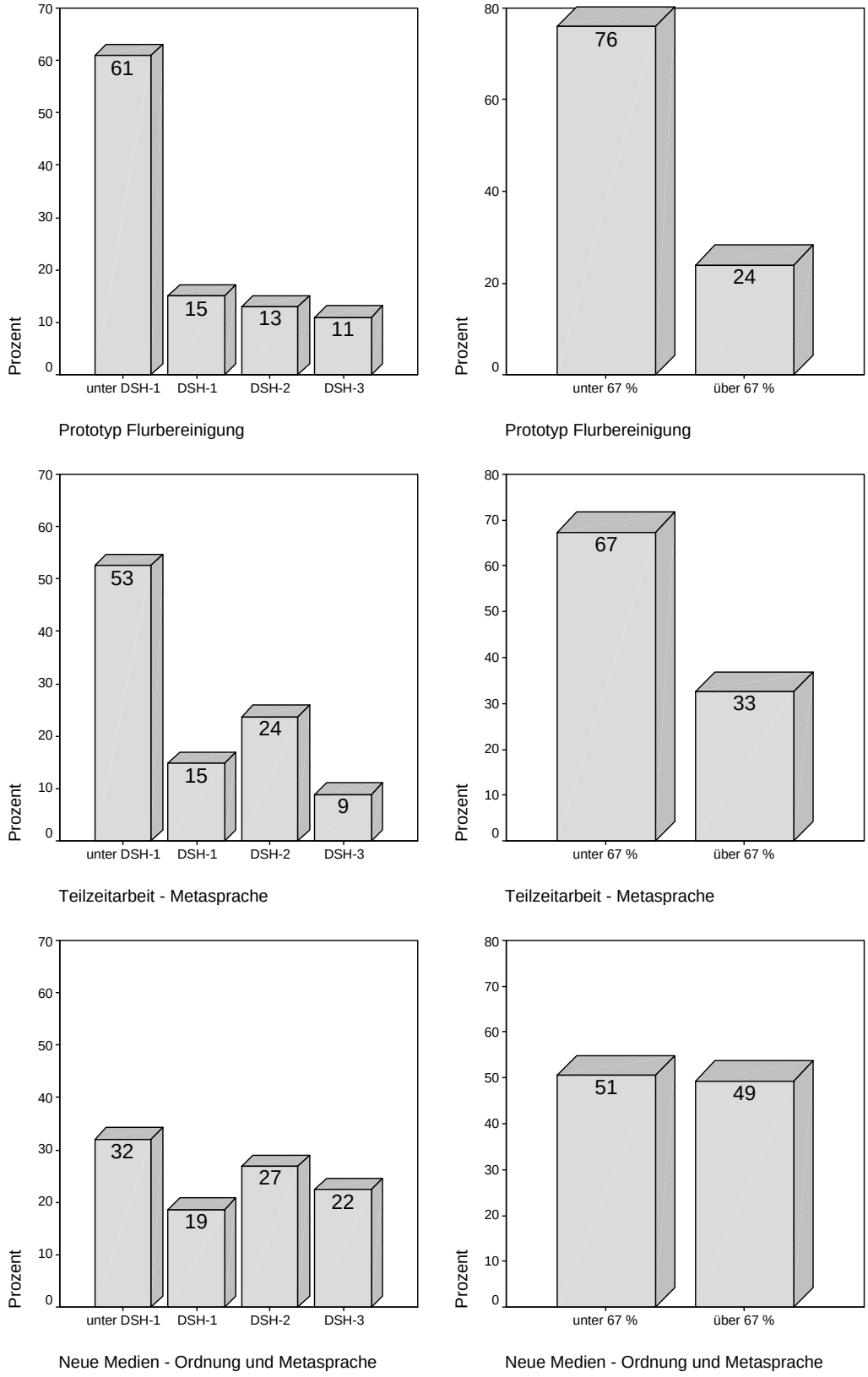
	Anzahl (n)	**Mittelwert** (*AM*) in %	***t*-Wert**	**Signifikanz** (p)
Prototyp "Flurbereinigung"	114	50 %	$t(113) = -2{,}858$	$p < 0{,}01$
"Teilzeitarbeit" – Metasprache	114	57 %		
"Neue Medien" – Ordnung	128	64 %	$t(127) = 4{,}983$	$p < 0{,}01$
"Teilzeitarbeit" – Metasprache	128	56 %		
Prototyp "Flurbereinigung"	135	51 %	$t(134) = -9{,}193$	$p < 0{,}01$
"Neue Medien" – Ordnung	135	65 %		

Tabelle 9: Grammatiktests – Korrelationen nach Pearson (r) und Übereinstimmungskoeffizienten

DSH-Grammatiktests	Korrelationskoeffizient nach Pearson (*r*)	Übereinstimmungskoeffizient, Schwellenwert bei 67 %*	Übereinstimmungskoeffizient, Schwellenwert nach Schwierigkeitsgrad**
Prototyp "Flurbereinigung" und Metasprache "Teilzeitarbeit" (n = 114)	**r = 0,437** $p < 0{,}01$	75,5 %	77,2 %
Metasprache "Teilzeitarbeit" und Ordnung "Neue Medien" (n = 128)	**r = 0,559** $p < 0{,}01$	66,4 %	77,3 %
Ordnung "Neue Medien" und Prototyp "Flurbereinigung" (n = 135)	**r = 0,681** $p < 0{,}01$	68,9 %	81,5 %

*Übereinstimmungskoeffizienten zeigen den Anteil der Kandidaten, die der gleichen Gruppe zugewiesen werden (Brown, 2001: 171). In dieser Spalte wird der Anteil der Kandidaten erfasst, die in beiden Tests weniger als 67 % oder in beiden Tests mehr als 67 % erzielten.

**In dieser Spalte wird der Übereinstimmungskoeffizient nach folgenden Schwellenwerten erfasst: "Flurbereinigung" > 67 %, "Teilzeitarbeit" > 72 %, "Neue Medien" > 80 %.



Anzahl (n): siehe Tabelle 7, Seite 106.

Abbildung 12: Ergebnisse in Grammatiktests nach Ergebnisklassen

Diskussion: Der Grammatiktest "Neue Medien", in dem die Items nach sprachlichen Phänomenen geordnet, mit Metasprache erläutert und mit einem Beispiel verstehen wurden, war signifikant leichter. Das ist nachvollziehbar, denn die Ordnung der Phänomene macht sie verständlicher. Der Grammatiktest "Teilzeitarbeit", in dem kurze metasprachliche Antworthinweise gegeben wurden, war mittelschwierig. Am schwierigsten war der Grammatiktest "Flurbereinigung" aus dem DSH-Handbuch, in dem die sprachlichen Phänomene ungeordnet und ohne metasprachliche Hinweise bearbeitet werden müssen. Dieses Ergebnis entspricht den Erwartungen, obwohl die Itemschwierigkeit sicherlich nicht allein von den Kriterien "Ordnung" und "Metasprache" abhängen, sondern auch von den Strukturen, die in den Items erfragt werden.

Eine Ordnung der Phänomene dürfte nicht nur zur Folge haben, dass der Test einfacher wird, sie führt in der Praxis auch dazu, dass die Einbindung in einen inhaltlichen Zusammenhang aufgegeben werden muss. Mit der Aufgabe der Kontextualisierung entfernt sich ein Grammatiktest noch weiter von einem "kommunikativen" oder "authentischen" Grammatiktest (siehe Kapitel 3.2, Seite 87).

Bemerkenswert sind die etwas niedrigeren Korrelationen (und damit die niedrige Paralleltestreliabilität) zwischen dem Grammatiktest mit Metasprache und den übrigen zwei Grammatiktests. Denkbar ist, dass die Verwendung von Metasprache einen Einfluss auf die Leistungen hat. Im Vergleich zum Test ohne Metasprache könnte dies bei einigen Kandidaten zu einer Ergebnisverbesserung führen, bei anderen zu einer Ergebnisverschlechterung, denn metasprachliche Hinweise können auch verwirren. Nur gemeinsam mit einer Ordnung der Phänomene ergab sich eine Verbesserung der Ergebnisse.

Deutlich wurde bei dem Einsatz von drei DSH-Grammatiktests mit unterschiedlichen Testmethoden-Merkmalen, dass allgemeine Angaben zu Schwellenwerten die Paralleltestreliabilität einschränken. Die Anweisung "DSH-2 ab 67 %" ist nicht sinnvoll, wenn die Tests unterschiedlich schwierig sind. Erstens weichen die Mittelwerte von DSH-Grammatiktests mit unterschiedlichen Testmethoden-Merkmalen signifikant voneinander ab. Die Chancen der Testteilnehmer sind größer, wenn sie eine DSH mit einem leichten Grammatiktest mit Ordnung und Metasprache bearbeiten müssen. Zweitens ist die Übereinstimmungsvalidität beeinträchtigt. In der Studie wiesen jeweils zwei Grammatiktests lediglich bei zwei Dritteln bis drei Vierteln aller Teilnehmer vergleichbare sprachlichen Leistungen aus. Es gibt keine allgemeine Regel, bei welchem Über-

einstimmungskoeffizienten man von einer zufrieden stellenden Reliabilität ausgeht. Auch wenn man bedenkt, dass die individuellen Leistungen von Test zu Test schwanken, sollte die Übereinstimmungsvalidität in diesem Fall meiner Ansicht nach deutlich höher liegen.

Wie verhält sich die Übereinstimmungsvalidität, wenn man die Schwellenwerte nach dem Schwierigkeitsgrad unterschiedlich festsetzt? Die Übereinstimmungskoeffizienten würden sich erhöhen (siehe Tabelle 9, Seite 106). Dies kann man an folgender, hypothetischer Vorgehensweise zeigen: Angenommen, die Testteilnehmer erfüllen das Testkriterium, wenn sie im Grammatiktest "Flurbereinigung" 67 Prozent erzielen. Dann wäre es schlüssig, wenn man den Schwellenwert beim Grammatiktest mit Metasprache "Teilzeitarbeit" bei 72 und beim Grammatiktest "Neue Medien" bei 80 Prozent ansetzen würde. Nur dann hätte die gleiche Anzahl der Teilnehmer das Testkriterium erfüllt, jeweils ungefähr 24 Prozent. Wie verhält es sich in diesem Fall mit der Übereinstimmungsvalidität? In diesem Fall steigt die Übereinstimmungsvalidität auf einen Wert zwischen 77 und 82 Prozent (siehe Tabelle 9, Seite 106), liegt also durchschnittlich über dem Wert, der für feste Schwellenwerte ermittelt wurde. Eine vergleichbare Veränderung ergibt sich, wenn die Schwellenwerte so festgelegt würden, dass beispielsweise jeweils 50 Prozent der Testteilnehmer das Testkriterium erfüllen.

Man muss davon ausgehen, dass die Testschwierigkeit von DSH-Grammatiktests keine feste Größe ist und nicht allein dem Methodeneffekt zuzuschreiben ist, sondern auch von der Schwierigkeit der einzelnen Items abhängt. Daher dürften feste Schwellenwerte auch dann nicht sinnvoll sein, wenn sicher gestellt wäre, dass der Methodeneffekt zu vernachlässigen ist.

Schließlich möchte ich noch auf Schwierigkeiten bei der Bewertung und Interpretation von Grammatiktests mit Metasprache hinweisen, die bei der Studie aufgefallen sind. Ein gewichtiges Argument für den Verzicht auf Metasprache ergibt sich aus der problematischen Auswertung derartiger Items. In einigen Fällen ist es möglich, dass es sprachlich korrekte Lösung gibt, die jedoch nicht im Sinne der Aufgabenstellung erfolgt ist. Dann stellt sich die Frage, wie die Antwort zu bewerten und zu interpretieren ist.

- Im Grammatiktest "Teilzeitarbeit" erhielt eine Testanweisung den Hinweis "Passiversatz". Dieser Hinweis wurde von einigen Testteilnehmern als Aufforderung

verstanden, einen Satz im Passiv zu konstruieren (passiver Satz vs. Passiv-ersatz!). Da der Ausgangssatz bereits im Passiv war und eine Änderung nicht möglich war, mussten sie bei dem Vorhaben, einen Satz im Passiv zu formulieren, scheitern, wenn sie den Satz nicht einfach wiederholen wollten. Bei dieser Aufgabenstellung erfährt man nicht, ob die Testteilnehmer mit dem Ausdruck "sein + zu + Infinitiv" ("..., dann ist die Alterssicherung zu reformieren") als Passiversatz vertraut waren, mit dem die Aufgabe zu lösen gewesen wäre. Man erfährt lediglich, dass der Begriff "Passiv-Ersatz", den man in der Tat leicht missverstehen kann, nicht bekannt war. Das ist meiner Ansicht nach aber eine belanglose Erkenntnis, die nicht mit dem Test-konstrukt zusammenhängt.

- Ein weiteres Item im Grammatiktest "Teilzeitarbeit" lautete: "... und dies ist für die meisten Männer noch nicht akzeptabel. (Passiv)" – Gefordert war die Lösung: "… und dies kann von den meisten Männern noch nicht akzeptiert werden." Wie soll man es bewerten und welche Schlüsse soll man ziehen, wenn einige Testteilnehmer als Lösung anbieten: "... und dies können die meisten Männer noch nicht akzeptie-ren."? Es handelt sich um eine sprachlich korrekte Lösung, die jedoch nicht in Ein-klang mit der Aufgabenstellung steht. Wiederum stellt sich das Problem der Bewer-tung und der Interpretation.

- Das letzte Beispiel für Probleme bei der Auswertung und Interpretation von Gram-matiktests mit Metasprache stammt aus dem Grammatiktest "Ordnung – Neue Medien". Ein Item enthielt die Anweisung: "Verkürzen Sie jeweils die Sätze, indem Sie die Nebensätze durch Satzglieder ersetzen." Ein zu verändernder Satz lautete: "Um den Einfluss ausländischer Medien zu kontrollieren, gibt es in einigen Staaten sehr strenge Gesetze." Als Lösung sah man wohl vor: "Zur Kontrolle des Einflusses..." Damit ist der Nebensatz durch ein Satzglied ersetzt worden und der Satz ist verkürzt worden. Die Antwort einiger Testteilnehmer lautete aber: "..., damit der Einfluss ausländischer Medien kontrolliert werden kann." – Das ist eine inhaltlich und sprachlich korrekte Lösung, die allerdings nicht der Testanweisung entspricht. Der Satz ist nicht kürzer geworden und der Nebensatz ist nicht durch ein Satzglied ersetzt worden.

4.1.3. Zusammenfassung und Diskussion

Fragestellung
Welche Auswirkungen haben unterschiedliche Testmethoden-Merkmale von Grammatiktests (mit/ohne Metasprache, mit/ohne Ordnung der sprachlichen Phänomene)?

In der Studie mit drei unterschiedlichen Grammatiktests wurde deutlich, dass unterschiedliche Testmethoden-Merkmale zu einem unterschiedlichen Schwierigkeitsgrad und zu unterschiedlichen Streuungen der Ergebnisse führen. Eine Ordnung der sprachlichen Phänomene und eine metasprachliche Erläuterung führte gemäß der Studie zu einem leichteren Test. Die Verwendung von Metasprache allein führte zu etwas veränderten Ergebnissen, je nachdem, ob sie mit einer Ordnung der Phänomene bzw. Beispielen einhergeht oder nicht. Die Paralleltestreliabilität von DSH-Grammatiktests mit unterschiedlichen Testmethoden-Merkmalen ist nicht als ausreichend hoch anzusehen.

Aus der Studie leite ich auch die Empfehlung ab, im DSH-Grammatiktest auf die Verwendung von Metasprache und auf eine Ordnung der sprachlichen Phänomene zu verzichten:

- Die Verwendung von Metasprache würde den Charakter der DSH als Kursabschlussprüfung verstärken. Es ist anzunehmen, dass Testteilnehmer von metasprachlichen Anweisungen profitieren, wenn sie vorher an einem Sprachkurs teilgenommen haben, in dem eben diese Terminologie auch eingeführt wurde. In einem Sprachtest für den Hochschulzugang sollte dies jedoch vermieden werden, so dass ein Verzicht auf Metasprache empfehlenswert ist.

- Bei der Auswertung der Tests mit Metasprache ist außerdem deutlich geworden, dass die Verwendung von Metasprache von dem Testkonstrukt ablenken kann. Das Testkonstrukt sollte die Kenntnis der wissenschaftssprachlichen Form und ihrer Bedeu-

tung sein. Die Frage, ob linguistische Terminologie verstanden wird, gehört nicht zum Testkonstrukt, ist nicht Konstruktrelevant (*construct irrelevance*).

- Eine Ordnung der sprachlichen Phänomene führt dazu, dass die Anbindung an einen Ausgangstext weitgehend aufgehoben werden muss. Dies ist ein Verlust an Authentizität.

Insgesamt empfiehlt sich also ein Verzicht auf eine Ordnung der sprachlichen Phänomene sowie auf metasprachliche Anweisungen. Dann müssen die Kandidaten bei jedem Item selbst erkennen, welche Transformation von ihnen verlangt wird. Auf diese Weise dürften sich die Ergebnisse im Grammatiktest bei gleichen sprachlichen Fähigkeiten unabhängig vom Besuch eines bestimmten prüfungsvorbereitenden Sprachkurses nicht unterscheiden. Zur Minimierung des Methodeneffekts sollten Hinweise zur Gestaltung des DSH-Grammatiktests auch in die Rahmenordnung aufgenommen werden. Dabei kann auf die Vorstellungen im DSH-Handbuch zurückgegriffen werden.

Als Empfehlung lässt sich weiter ein Verzicht auf allgemeine Angaben zum kritischen Wert ableiten. Wenn unterschiedliche Testmethoden-Merkmale verwendet werden, ist eine Anweisung "DSH-2 ab 67 %" in der DSH-Rahmenordnung nicht sinnvoll. Da die DSH keine standardisierte Prüfung ist, sollte auf derartige Anweisungen verzichtet werden. Auch wenn sicher gestellt sein sollte, dass die DSH durchgängig mit gleichen Testmethoden-Merkmalen erstellt wird, muss der kritische Wert in Abhängigkeit vom Schwierigkeitsgrad festgesetzt werden. Unabdingbar ist daher eine inhaltliche Beschreibung der geforderten Fertigkeiten, die über die kurzen Hinweise auf der Rückseite der Zeugnisse hinausgeht.

Zur Legitimität des DSH-Grammatiktests ist festzuhalten, dass eine Einheitlichkeit der Testmethoden-Merkmale eine unbedingte Voraussetzung für die Reliabilität darstellt. Wenn diese nicht sichergestellt werden kann, dürfte die Vergleichbarkeit der DSH-Prüfungen durch den DSH-Grammatiktest nur noch verringert werden. Mit der Registrierung von DSH-Prüfungen unternehmen Hochschulrektorenkonferenz und Fachverband Deutsch als Fremdsprache einen Schritt in die richtige Richtung. Dieser geht allerdings noch nicht weit genug: Es sollten bei der Registrierung Musterprüfungen verlangt werden, in denen die Testmethoden-Merkmale überprüft werden. Wichtig sind auch

Schulungen von Prüfern. Letztlich muss man aber eingestehen, dass die Reliabilität bei nicht standardisierten Prüfungen immer auf einem niedrigen Niveau sein wird.

4.2. Konstruktvalidität des DSH-Grammatiktests

Übersicht: Kapitel 4.2

Welche legitimen Interpretationen Ergebnisse im DSH-Grammatiktest ermöglichen, ist Thema des folgenden Kapitels. Ich greife auf drei unterschiedliche Untersuchungen zurück: auf eine Studie mit muttersprachlich deutschen Studierenden, auf Ergebnisse aus der DSH-TestDaF-Pilotstudie sowie auf Ergebnisse aus der DSH-TestDaF-Vergleichsstudie. Informationen zur Konstruktvalidität werden mithilfe von statistischen Verfahren erhoben: Korrelationstests, Regressionsanalysen und Faktoranalysen.

4.2.1. Fragestellung und Methode

Von der Frage nach der Äquivalenz von Grammatiktests komme ich zum Problem der Validität: Was misst ein Grammatiktest? Wie können die Leistungen in Grammatiktests interpretiert werden? In Kapitel 2.2 (Seite 49 ff) wurde die Diskussion um die Konstruktvalidität bereits beschrieben und auf die Bestandteile der Konstruktvalidität nach Messick und Bachman/Palmer verwiesen. In Kapitel 3.2 (Seite 82 ff) wurde darauf hingewiesen, dass es wünschenswert wäre, weitere Aussagen zur Konstruktvalidität von Grammatiktests im Zusammenhang mit Sprachstandstests für den Hochschulzugang zu gewinnen. Die theoretische Bestimmung des Konstrukts von Grammatiktests lautet: produktive Grammatikkompetenz (siehe Kapitel 3.2, Seite 82 ff). Von Interesse ist, zu erfahren, ob Grammatiktests mehr erfassen als die Fähigkeit, sprachliche Transformationsaufgaben auszuführen.

Fragestellung

Der DSH-Grammatiktest misst produktive Grammatikkompetenz. Welche weiteren Interpretationen des Testergebnisses sind angemessen?

Um Aussagen zur Konstruktvalidität des DSH-Grammatiktests zu gewinnen, nutze ich Daten, welche im Rahmen unterschiedlicher Erhebungen gewonnen wurden: die DSH-TestDaF-Pilotstudie an der FH Konstanz, die DSH-TestDaF-Vergleichsstudie sowie eine Studie mit deutschen Studierenden. Diese drei Studien werden im Folgenden beschrieben und die Vorgehensweise wird begründet. Die Ergebnisse der drei Studien werden in den Kapiteln 4.2.2, 4.2.3 bzw. 4.2.4 dargestellt, Zusammenfassung und Diskussion in Kapitel 4.2.5.

Studie mit muttersprachlich deutschen Studierenden

Stellen die im DSH-Grammatiktest geprüften sprachlichen Fertigkeiten besonders fremdsprachige Deutschlerner vor Schwierigkeiten oder stellen die Anforderungen auch Muttersprachler vor Schwierigkeiten? Ein Vergleich der Leistungen von Muttersprachlern und Deutschlernern soll Hinweise auf die Konstruktvalidität des DSH-Grammatiktests ermöglichen. Ein Argument, das häufig in einem informellen Rahmen gegenüber dem DSH-Grammatiktest angebracht wird, lautet: Auch deutsche Studenten hätten große Schwierigkeiten mit den im DSH-Grammatiktest geforderten sprachlichen Transformationen und könnten diese ohne eine gezielte Vorbereitung ebenfalls nicht lösen. Sollte das zutreffen, stellt sich in der Tat die Frage, warum ausländische Studienbewerber mit derartigen Aufgaben konfrontiert werden. Sollten muttersprachlich deutsche Studierende Schwierigkeiten mit dem DSH-Grammatiktest haben, so ließen sich zusätzliche Hinweise auf die Konstruktvalidität gewinnen.

Ob muttersprachlich deutsche Studierende Schwierigkeiten mit dem DSH-Grammatiktest haben, wurde in einer gesonderten Studie im Rahmen eines Erstsemesterseminars an der Fachhochschule Konstanz erhoben. Bei den Seminarteilnehmern handelte es sich um 76 deutsche und zusätzlich um einige ausländische Studierende der Fachrichtungen Elektrotechnik und Projektingenieurwesen. Die Tests der Studierenden nichtdeutscher Muttersprache wurden nicht berücksichtigt. Ich legte den Teilnehmern den DSH-Prototyp-Grammatiktest "Flurbereinigung" aus dem DSH-Handbuch vor (siehe Abbildung 7, Seite 74). Die Leistungen der muttersprachlich deutschen Studierenden konnten mit denen der ausländischen Studierenden verglichen werden, welche in

anderen Studien erhoben wurden. Die Ergebnisse dieser Studie werden in Kapitel 4.2.2 (Seite 127 ff) vorgestellt und diskutiert.

DSH-TestDaF-Pilotstudie an der FH-Konstanz

Mit Teilnehmern aus dem Studienkolleg der Fachhochschule Konstanz wurde die "DSH-TestDaF-Pilotstudie" durchgeführt. In dieser Studie wurden die Ergebnisse aus dem TestDaF und dem DSH-Grammatiktest verglichen. Ein Ziel der Studie war es, eine Empfehlung für die Bewertung von TestDaF-Zeugnissen zu erhalten. Die Ergebnisse der Studie wurden bereits veröffentlicht (Krekeler, 2002b). In der vorliegenden Arbeit geht es jedoch nur um Informationen zur Konstruktvalidität des DSH-Grammatiktests. Im Rahmen der Studie nahmen 67 Kollegiaten an einer Erprobungsfassung des TestDaF teil, welche auch einen C-Test umfasste (siehe Seite 8). Außerdem bearbeiteten sie den Grammatiktest "Flurbereinigung" (siehe Abbildung 7, Seite 74).

Der TestDaF, welcher von den Studierenden bearbeitet wurde, war eine Erprobungsfassung in der zweiten Stufe. In dieser Phase werden die vorerprobten Aufgaben nochmals auf ihre Güte überprüft, und die Ergebnisse der Prüfungsteile Leseverstehen sowie Hörverstehen werden den TestDaF-Niveaustufen zugeordnet. Dieser Prozess der Kalibrierung erfolgt mittels mehrerer C-Tests, deren Schwierigkeitsgrade bekannt sind (Arras/Eckes/Grotjahn, 2002; Arras/Grotjahn, 2002; TestDaF-Institut, 2001). Die eingesetzte Erprobungsfassung unterschied sich nur geringfügig von der Endversion. Der DSH-Grammatiktest "Flurbereinigung" wurde von mir korrigiert, der TestDaF von den Korrekteuren des TestDaF-Instituts.

Die Stichprobe entspricht in wesentlichen Merkmalen der Population "ausländische Studienbewerber". Die Verteilung der Herkunftsländer und Ausgangssprachen war beispielsweise repräsentativ für ausländische Studierende (im Vergleich mit den Zahlen aus DAAD/HIS, 2004). Hinsichtlich der Herkunft und der Muttersprache(n) waren die Gruppen sehr heterogen. Die größten Gruppen stellten Studierende aus China, Marokko und dem Libanon. Die Stichprobe ist mit einer Einschränkung als spezifisch repräsentativ anzusehen. Die Teilnehmer der Pilotstudie besuchten das Studienkolleg der Fachhochschule Konstanz. Nicht alle ausländischen Studienbewerber haben die Möglichkeit bzw. die Pflicht, an einem einjährigen studienvorbereitenden Kurs teilzunehmen. Die

Studierenden werden in zwei Semestern auf die DSH, die sie als Teil der Feststellungsprüfung ablegen, vorbereitet.

Tabelle 10: DSH-TestDaF-Pilotstudie an der Fachhochschule Konstanz – Prüfungsteile

TestDaF Mündlicher Ausdruck (MA), Schriftlicher Ausdruck (SA) Die TestDaF-Prüfungsteile Mündlicher Ausdruck und Schriftlicher Ausdruck führen zu den folgenden Ergebnisklassen: "unter TDN 3", TDN 3, TDN 4 bzw. TDN 5.
TestDaF Hörverstehen und Leseverstehen (HV, LV) Beim Hörverstehen sind 25 Items zu lösen, beim Leseverstehen insgesamt 30. Für jedes richtig gelöste Item gibt es einen Punkt. Die TestDaF-Niveaustufen werden aus den Rohwerten ermittelt.
C-Test Der C-Test besteht aus vier Texten mit insgesamt 80 Lücken. Nur lexikalisch, orthografisch und grammatisch korrekte Lösungen werden als richtig gewertet. Für jede richtige Lösung gibt es einen Punkt. Das TestDaF-Institut in Hagen verwendet bei der Auswertung des C-Tests umfangreichere Auswertungskategorien (Arras/Eckes/Grotjahn, 2002; Hinweise zu C-Tests siehe Seite 8).
Wissenschaftssprachliche Strukturen/DSH-Grammatiktest "Flurbereinigung" (siehe Abbildung 7, Seite 74) Die Bewertung des ersten DSH-Grammatiktests erfolgte anhand der im "DSH-Handbuch für Prüferinnen und Prüfer" beschriebenen Punkteverteilung (FaDaF, 2001: 7/6). Er bestand aus zehn Sätzen (Items). Es konnten maximal 26,5 Punkte erzielt werden. Das kleinste Intervall bestand aus 0,5 Punkten. Die Rohwerte wurden in eine Notenskala (1,0 bis 5,0) übertragen.

Die Studierenden werden im Studienkolleg der Fachhochschule Konstanz in vier Gruppen mit 18 bis 24 Teilnehmern unterrichtet. Es gibt zwei Lerngruppen für Teilnehmer, die ein technisches Studienfach anstreben (auch Informatik oder Architektur), und zwei Lerngruppen für Teilnehmer, die wirtschaftswissenschaftliche Studiengänge belegen möchten. Dass sich die sprachlichen Leistungen der beiden Gruppen unterscheiden, ging aus den Ergebnissen des vom TestDaF-Institut entwickelten C-Tests hervor, welcher zur Erprobungsfassung des TestDaF gehört. Die Leistungen der Studierenden aus den W-Kursen lagen deutlich über denjenigen aus T-Kursen (W-Kurs: $AM = 60$ %, T-Kurs: $AM = 51$ %). Wenn man davon ausgeht, dass der C-Test ein gutes Maß für die allgemeine Sprachkompetenz darstellt, ist die Sprachkompetenz der Kollegiaten aus dem W-Kurs signifikant höher als diejenige der Kollegiaten aus dem T-Kurs (t-Test für unabhängige Stichproben: $t(62) = 2{,}487$, $p < 0{,}05$ 2-seitig). Die Deutschkenntnisse der Studienbewerber mit den Studienzielen Informatik, Maschinenbau, Elektrotechnik usw. liegen demnach deutlich unter derjenigen von Studienbewerbern, welche wirtschaftswissenschaftliche Studiengänge anstreben. Dies ist mit Blick auf eine differenzierte

Zulassungspraxis, welche der TestDaF ermöglicht, zwar interessant, aber kein Hinweis darauf, ob Studierende technischer Fächer weniger Deutschkenntnisse benötigen.

An der Pilotstudie nahmen Studierende aus dem ersten und dem zweiten Semester des Studienkollegs teil. Wie zu erwarten, lagen die Ergebnisse der Studierenden aus dem zweiten Semester im C-Test im Mittel über denen aus dem ersten Semester lagen (1. Sem.: *AM* = 51,4 %; 2. Sem.: *AM* = 59,6 %). Auch dieser Unterschied ist signifikant ($t(62) = 2{,}17$, $p < 0{,}05$ 2-seitig). Die Studierenden aus dem ersten Semester hatten einige Wochen vorher an einem von uns konzipierten Aufnahmetest teilgenommen, der auch einen Deutschteil enthält. Die andere Gruppe der Teilnehmer befand sich am Beginn des zweiten Semesters. Sie hatten bereits ein Semester des Studienkollegs absolviert und wurden in das zweite Semester versetzt. Voraussetzung dafür sind u. a. mindestens ausreichende Leistungen im Fach Deutsch.

Die Ergebnisse der DSH-TestDaF-Pilotstudie werden in Kapitel 4.2.3 (Seite 129 ff) vorgestellt und diskutiert.

DSH-TestDaF-Vergleichsstudie

Im Rahmen der DSH-TestDaF-Vergleichsstudie nahmen 56 Probanden in Konstanz sowohl am TestDaF als auch an der gesamten DSH teil. Daher können die Ergebnisse des DSH-Grammatiktests mit den übrigen Prüfungsteilen der DSH und des TestDaF in Verbindung gebracht werden.

Ziel der DSH-TestDaF-Vergleichsstudie war es, Hinweise für eine Zulassungspraxis mit der DSH und dem TestDaF auf einer breiten Basis zu gewinnen. Dazu nahmen ausländische Studienbewerber an vier Hochschulen sowohl an der DSH als auch am TestDaF teil. Vorläufige Ergebnisse der DSH-TestDaF-Vergleichsstudie wurden bereits veröffentlicht (Koreik, 2003; 2005. Hier sollen die Eckdaten der DSH-TestDaF-Vergleichsstudie erwähnt werden, die für die vorliegenden Studien relevant sind. Hier werden nur Daten verwendet, die an der Fachhochschule Konstanz erhoben wurden. Es nahmen 56 Kandidaten zunächst an der DSH und eine Woche später am TestDaF teil (siehe Tabelle 11). Zusätzlich füllten die Teilnehmer einen umfangreichen Fragebogen zu ihrer Sprachlernbiografie aus. Die Teilnahme an der Vergleichsstudie war kostenlos

und freiwillig. Als die Probanden am TestDaF teilnahmen, waren ihnen die Ergebnisse aus der DSH bereits bekannt.

Tabelle 11: DSH-TestDaF-Vergleichsstudie – Ablauf an der FH Konstanz

Phase 1: DSH
Prüfungsteil: DSH-Hörverstehen
Prüfungsteil: DSH-Leseverstehen (Zwei Texte zur Auswahl)
Prüfungsteil: Wissenschaftssprachliche Strukturen/DSH-Grammatiktest "Meinungs-forschung" (siehe Abbildung 13, Seite 124)
Prüfungsteil: DSH-Textproduktion (mehrere Themen zur Auswahl)
Prüfungsteil: DSH-Mündliche Prüfung (nicht alle Kandidaten)
Fragebogen und Interviews
Ausführlicher Fragebogen zur Sprachlernbiografie (bei der DSH ausgeteilt und vor dem TestDaF eingesammelt)
Semistrukturierte Interviews
Phase 2: TestDaF – 1 Woche später
Prüfungsteil: TestDaF-Hörverstehen
Prüfungsteil: TestDaF-Leseverstehen
Prüfungsteil: TestDaF-Schriftlicher Ausdruck
Prüfungsteil: TestDaF-Mündlicher Ausdruck

Bei den 56 Testteilnehmern am Standort Konstanz handelte es sich um ausländische Studienbewerber, die ein Studium an einer Fachhochschule in Baden-Württemberg anstrebten. Nach ihrem Status lassen sich vier Gruppen unterscheiden:

- **Die Kollegiaten**: Vier Testteilnehmer verfügten nicht über eine Direktzulassung zum Studium an einer deutschen Hochschule; sie mussten sich durch den Besuch des Studienkollegs dazu qualifizieren. Zum Testzeitpunkt hatten sie das erste Semester des Studienkollegs bereits besucht. Sie hatten von der Möglichkeit zur Teilnahme an den Prüfungen erfahren und nahmen an der DSH und am TestDaF teil, um ihren Sprachstand besser einschätzen zu können.

- **Die Austauschstudentin**: Eine Teilnehmerin war als Austauschstudentin ein Semester in Deutschland. Sie war als eine der Klassenbesten ihres Germanistik-

Studiengangs in China für das Austauschprogramm ausgewählt worden. Sie nahm an der DSH teil, weil sie nach dem Abschluss ihres Studiums in China in Deutschland weiter studieren wollte. Am TestDaF nahm sie zusätzlich aus Interesse teil.

- **Die DSH-Kurs-Teilnehmer**: Zehn Testteilnehmer hatten an einem dreimonatigen Deutschkurs an der Fachhochschule Konstanz zur Vorbereitung auf die DSH teilgenommen. Sie verfügten über eine Direktzulassung. Sie nahmen zusätzlich am TestDaF teil, weil sie sich für den TestDaF interessierten oder weil sie damit rechneten, die DSH möglicherweise nicht zu bestehen.

- **Die "Unbekannten" mit Direktzulassung**: Die größte Gruppe der Testteilnehmer war uns nicht durch das Studienkolleg oder durch studienvorbereitende Sprachkurse bekannt. Sie hatten Kontakt zum Studienkolleg der Fachhochschule Konstanz, weil ihre Zeugnisse im Auftrag anderer Fachhochschulen geprüft wurden. Die 41 Kandidaten hatten sich allein oder in Sprachkursen vorbereitet, die vor allem von außeruniversitären Anbietern ausgerichtet wurden. Sie haben die Gelegenheit zur kostenlosen Teilnahme an einer weiteren Sprachprüfung aus Interesse wahrgenommen oder weil sie ebenfalls Bedenken hatten, ob sie die DSH bestehen würden. Mit der Teilnahme am TestDaF wollten sie ihre Chancen erhöhen.

Die Gruppe der Probanden sollte repräsentativ für die Gruppe der ausländischen Studienbewerber an der Fachhochschule Konstanz sein. Mit Blick auf die Kriterien Studienziele, Herkunftsländer, Muttersprachen, Alter, Aufenthaltsdauer in Deutschland und Verteilung der Geschlechter ist eine repräsentative Auswahl der Probanden durchaus gelungen. Nicht repräsentativ war die Probandengruppe mit Blick auf das Niveau der Deutschkenntnisse.

- **Studienziele**: Die meisten Testteilnehmer gaben an, entweder ein betriebswirtschaftliches oder ein technisches Studium anzustreben. Achtmal wurde Informatik bzw. Wirtschaftsinformatik als Studienziel genannt. Nur einzelne Testteilnehmer wollten Architektur, Design oder Sozialwesen studieren.

- **Herkunftsländer und Muttersprachen**: Die größte Gruppe stellten die 14 Studienbewerber aus China. Jeweils fünf Testteilnehmer stammten aus Kamerun, aus der Türkei bzw. aus Tunesien. Jeweils drei Testteilnehmer kamen aus Estland, Indonesien oder Vietnam. Außerdem waren Studienbewerber aus 15 weiteren Ländern ver-

treten. Diese Verteilung spiegelt sich auch in den Herkunftssprachen: 15 Testteilnehmer sprechen Chinesisch als Muttersprache, je fünf Französisch oder Türkisch, je vier Russisch oder Arabisch. Damit ist die Probandengruppe repräsentativ für DSH-Teilnehmer an der Fachhochschule Konstanz. Allein die Gruppe der Chinesen ist überrepräsentiert.

- **Alter und Aufenthaltsdauer in Deutschland**: Auch was das Alter betrifft, dürfte die Gesamtgruppe der ausländischen Studienbewerber ungefähr erfasst worden sein. Die Testteilnehmer waren zwischen 18 und 36 Jahren alt. Das durchschnittliche Alter betrug knapp 25 Jahre. Die meisten Testteilnehmer hatten sich bislang zwischen 4 Monaten und 3 Jahren in Deutschland aufgehalten.

- **Verteilung der Geschlechter**: Die Verteilung der Geschlechter war ungefähr gleich. Es nahmen 27 Frauen und 29 Männer teil. Die Mittelwerte der Ergebnisse beider Gruppen in der DSH unterscheiden sich nicht signifikant, wohl aber die Ergebnisse im TestDaF: Das Ergebnis der Testteilnehmerinnen lag signifikant über dem Ergebnis der Testteilnehmer, wie mit einem *t*-Test für unabhängige Stichproben gezeigt werden kann (siehe Tabelle 12, Seite 123). Man könnte erwarten, dass sich die Sprachkenntnisse der beiden Gruppen entweder unterscheiden oder nicht. Aber dass die DSH die Deutschkenntnisse auf vergleichbarem Niveau ausweist, der TestDaF jedoch nicht, ist überraschend. Über die Gründe, die zum schlechteren Abschneiden der Männer im TestDaF bzw. zum besseren in der DSH führten, kann nur spekuliert werden. Möglicherweise waren die Männer beim TestDaF weniger motiviert, strengten sich nicht mehr an, nachdem sie die Ergebnisse aus der DSH bereits kannten. Die Frauen wären – dieser Argumentation folgend – verlässlicher.

- **Niveau der Deutschkenntnisse**: Informationen zum Niveau der Deutschkenntnisse der Probanden konnten durch einen Vergleich gewonnen werden: An der DSH nahmen nämlich nicht nur die 56 Probanden der DSH-TestDaF-Vergleichsstudie, sondern auch weitere 180 ausländische Studienbewerber teil. Die statistischen Kennwerte weisen darauf hin, dass das Niveau der Deutschkenntnisse der Probanden niedriger war als das der Gesamtgruppe (siehe Tabelle 13, Seite 123): Die Teilnehmer an der Vergleichsstudie erzielten in der DSH im Durchschnitt die Note 3,6. Die übrigen Teilnehmer erzielten in der DSH ein besseres Ergebnis, es lag durchschnittlich bei der Note 3,3 (auf einer Skala von 1,0 bis 5,0). Der Unterschied zwischen den

beiden Mittelwerten ist auf dem Niveau 95 Prozent signifikant, wie man mit einem *t*-Test für unabhängige Stichproben zeigen kann. Vor allem Kandidaten mit sehr fortgeschrittenen Deutschkenntnissen verzichteten offensichtlich auf eine Teilnahme an der Vergleichsstudie (d. h. zusätzlich am TestDaF). Während die beste der 56 Teilnehmerinnen und Teilnehmer der Vergleichsstudie in der DSH die Note 1,8 erzielte, erreichte die beste der übrigen 180 die Note 1,1. Insgesamt kann man feststellen, dass über geringere Deutschkenntnisse verfügte, wer sich zur Teilnahme an der DSH-TestDaF-Vergleichsstudie entschied. Aus der Vorgehensweise bei der Auswahl der Probanden liegen folgende Annahmen nahe: Erstens dürften die Kandidaten vor allem auf die DSH vorbereitet gewesen sein, da sie erst später von der Möglichkeit erfuhren, zusätzlich am TestDaF teilzunehmen. Abgesehen von der Information über das Angebot zur zusätzlichen Teilnahme an einer weiteren Deutschprüfung wurden die ausländischen Studienbewerber nicht über das Format des TestDaF informiert. Zweitens war die Teilnahme am TestDaF vor allem für Kandidaten von Interesse, die damit rechnen mussten, die DSH nicht zu bestehen. Wenn sie tatsächlich nicht bestanden hatten, war die Teilnahme am TestDaF umso interessanter. Es dürfte sich also um Kandidaten gehandelt haben, bei denen das Niveau der Deutschkenntnisse möglicherweise eher gering war.

Die eingesetzte DSH wurde nach den Richtlinien aus dem DSH-Handbuch vor Ort erstellt. Der DSH-Grammatiktest "Wissenschaftssprachliche Strukturen" bestand aus zehn Items (siehe Abbildung 13). Bei jedem Item waren maximal drei Punkte zu erzielen, bei einem Item lediglich zwei, so dass sich die maximal erreichbare Punktzahl auf 29 belief. Der Schwellenwert wurde bei 14,5 Punkten festgelegt. Die Aufgabe lautete: "Füllen Sie die Lücken aus, ohne die Textinformationen zu verändern! Die Unterstreichungen sollen Ihnen bei der Lösung helfen." Verlangt wurden folgende Umformungen: Modalverb in ein modalverbähnliches Verb (2x), Verbalstil in Nominalstil mit Funktionsverbgefüge (2x), Partizip als Attribut in einen Relativsatz (2x), Relativsatz in ein Partizip als Attribut, Präpositionaler Ausdruck in einen Nebensatz, passiver Ausdruck in einen aktiven Ausdruck, Infinitivsatz in einen Nebensatz. Der Text, welcher den Transformationsitems zugrunde lag, handelte von der Meinungsforschung.

Tabelle 12: DSH-TestDaF-Vergleichsstudie: Ergebnisse der Frauen und der Männer im Vergleich

	Anzahl (n)	**Mittleres Ergebnis in der DSH (AM)***	**Mittleres Ergebnis im TestDaF (AM)****
Weiblich	27	3,49	3,56
Männlich	29	3,43	2,96
Signifikanz		t(54) = -0,199; nicht signifikant	t(54) = -3,732; p < 0,01

* *AM* (Arithmetisches Mittel) der Zensur, 1,0 – 5,0.
** *AM* der TestDaF-Niveaustufen ("unter TDN 3" als "2" codiert)

Tabelle 13: Teilnehmer an der DSH-TestDaF-Vergleichsstudie und übrige Teilnehmer an der DSH – statistische Kennwerte

	Nur Teilnahme an der DSH	**Zusätzlich Teilnahme am TestDaF**	**Gesamt**
Anzahl (n)	180	56	236
Mittelwert (Arithmetisches Mittel; *AM*; Skala: 1,0-5,0)	3,319	3,643	3,396
Vergleich der Mittelwerte (*t*-Test für unabh. Stichproben)	$t_{(234)} = -2{,}403$; $p = 0{,}017$		
Median (Wert, der die Gesamtzahl in zwei Hälften teilt; *Md*; Skala: 1,0-5,0)	3,311	3,640	3,460
Standardabweichung (s)	0,917	0,754	0,890
Minimum – Maximum	1,1 – 5,0	1,8 – 5,0	1,1 – 5,0

Füllen Sie die Lücken aus, ohne die Textinformation zu verändern! Die Unterstreichungen sollen Ihnen bei der Lösung helfen.	
Die Meinungsforschung (Demoskopie)	
Die Meinungsforschung (Demoskopie) ist ein wissenschaftliches Verfahren, mit dem die Meinung der Bevölkerung erforscht wird. Die Meinungsforschung erlebte ihren Durchbruch 1936 in den Vereinigten Staaten, als George Gallup auf der Grundlage einer repräsentativen Stichprobe den Ausgang der amerikanischen Präsidentschaftswahlen richtig vorhersagte - anders als die Zeitschrift "Literary Digest".	
Beispiel: Die Redakteure der Zeitschrift "Literary Digest" waren nach Auswertung von über zwei Millionen Fragebögen davon überzeugt, dass Alfred M. Landon und nicht Franklin D. Roosevelt die Wahl gewinnen werde.	Nachdem die Redakteure der Zeitschrift "Literary Digest" über zwei Millionen Fragebögen ausgewertet hatten, waren sie davon überzeugt, dass Alfred M. Landon und nicht Franklin D. Roosevelt die Wahl gewinnen werde.
1. Gallup befragte nur wenige tausend Personen. Dadurch konnte er nicht nur das Wahlergebnis, sondern auch die zu erwartende Fehlschätzung des "Literary Digest" prognostizieren.	Gallup konnte durch . . . . nicht nur das Wahlergebnis, sondern auch die zu erwartende Fehlschätzung des "Literary Digest" prognostizieren.
2. Plötzlich wurde sich die Öffentlichkeit bewusst, dass man die Wahrscheinlich-keitsrechnung auf die politische Meinungsbildung anwenden konnte.	Plötzlich wurde sich die Öffentlichkeit bewusst, dass sich. . . . .
Mit diesem spektakulären Erfolg setzte der Siegeszug der Demoskopie ein.	Mit diesem spektakulären Erfolg setzte der Siegeszug der Demoskopie ein.
3. Der amerikanische Präsident Franklin D. Roosevelt beispielsweise ließ sich ab Beginn der vierziger Jahre von Hadley Cantril, einem an der Universität Princeton lehrenden Meinungsforscher, beraten.	Der amerikanische Präsident Franklin D. Roosevelt beispiels-weise ließ sich ab Beginn der vierziger Jahre von Hadley Cantril, einem Meinungsforscher, ... ,beraten.
[...]	

(vollständiger Test: Anhang 4, Seite 360)

Abbildung 13: DSH-Grammatiktest "Meinungsforschung" aus der DSH-TestDaF-Vergleichsstudie

Statistische Verfahren zur Konstruktvalidität

In diesem Teil der Studie setze ich mehrere statistische Verfahren ein, mit denen Aussagen zur Konstruktvalidität gewonnen werden sollen.

Mit **Korrelationsanalysen** wird die Stärke der Beziehungszusammenhänge zwischen dem DSH-Grammatiktest und anderen Prüfungsteilen überprüft. Die Korrelation, das Maß der Übereinstimmung zwischen zwei Messwerten, lässt sich mit Korrelationskoeffizienten bestimmen (Bachman, 2004). Ein Korrelationskoeffizient liegt zwischen -1 und +1. Je höher die Korrelation, desto näher liegt der Wert an + oder -1. Eine Korrelation ist nicht unbedingt ein Hinweis auf einen ursächlichen Zusammenhang. Wenn der Grammatiktest jedoch *nicht* mit Tests korreliert, deren Konstrukt bekannt ist, muss man davon ausgehen, dass *keine* Übereinstimmungsvalidität (*concurrent validity*) vorliegt.

Ein weiteres statistisches Verfahren, mit dem sich Argumente für die Konstruktvalidität sammeln lassen, ist die **Faktoranalyse**. Eine Faktoranalyse ist ein datenreduzierendes Verfahren zur Entdeckung von Zusammenhängen. Es kann geprüft werden, ob Merkmale auf einige wenige "zentrale Faktoren" zurückgeführt werden können (Bortz/Döring, 2002; Backhaus/Erichson/Plinke/Weiber, 2003). Mit Hilfe der Faktoranalyse sollten zentrale Faktoren entdeckt und beschrieben werden, die den Ergebnissen zugrunde liegen. An Sprachtests für den Hochschulzugang kann mittels einer Faktoranalyse erhoben werden, welche Prüfungsteile Ähnliches messen und sich möglicherweise auf ein Konstrukt zurückführen lassen. Eine "Verdichtung" der von mehreren Prüfungsteilen gemessenen Eigenschaften kann auch Hinweise darauf geben, ob ein Grammatiktest ein sinnvoller Bestandteil eines Sprachtests für den Hochschulzugang ist.

Während es sich bei der Faktoranalyse um ein Struktur-enteckendes Verfahren handelt, ist die **Regressionsanalyse** ein Struktur-prüfendes Verfahren. Mit Regressionsanalysen können Zusammenhänge zwischen mehreren Variablen beschrieben und erklärt werden. Es können auch Prognosen über die Wirkungsbeziehungen aufgestellt werden. Mit Regressionsanalysen wurde untersucht, ob und in welchem Umfang die Ergebnisse des Grammatiktests durch andere Prüfungsteile vorhergesagt bzw. erklärt werden können. Dabei wurde der DSH-Grammatiktest als Kriteriumsvariable gewertet, die mittels eines

linearen Gleichungsmodells auf der Basis anderer Prüfungsteile, welche als Prädiktorvariablen eingesetzt werden, vorhergesagt werden sollte (Bortz/Döring, 2002; Backhaus/Erichson/Plinke/Weiber, 2003).

Einige Anmerkungen zu den **Skalenniveaus**: Mit Ausnahme der TestDaF-Niveaustufen (TDN) liegen bei den Ergebnissen der Tests (Prüfungsteile der DSH und C-Tests) intervallskalierte Daten vor, d. h. die Abstände zwischen den einzelnen Skalenwerten sind konstant. Bei den TestDaF-Niveaustufen könnte das Vorliegen einer Intervallskala bezweifelt werden, was an dem Ergebnis "unter TDN 3" liegt. Beim TestDaF werden alle Ergebnisse unter der TDN 3 der Ergebnisklasse "unter TDN 3" zugeordnet. Ein Kandidat, der überhaupt keine Deutschkenntnisse hat, würde das gleiche Ergebnis erzielen, wie jemand, der über Deutschkenntnisse auf Grundstufenniveau verfügt. Im Rahmen dieser Studie dürfte dieser Fall jedoch nicht aufgetreten sein. Da die Kollegiaten ihre fortgeschrittenen Deutschkenntnisse bereits im Aufnahmetest unter Beweis gestellt haben, dürfte sich das Ergebnis "unter TDN 3" wie eine weiter Ergebnisklasse auf einer intervallskalierten Skala verhalten. Daher wurden auch bei den TestDaF-Ergebnissen parametrische Verfahren angewandt, welche intervallskalierte Daten voraussetzen (Regressions- und Faktoranalysen, Korrelationskoeffizient nach Pearson). Zur Kontrolle wurden bivariate Korrelationen häufig auch mit Hilfe von Rangkorrelationskoeffizienten bestimmt, welche von der "niedrigeren" Rang- oder Ordinalskala ausgehen.

4.2.2. Ergebnisse der Muttersprachler im DSH-Grammatiktest und Diskussion

Ergebnisse: Die deutschen Studenten erzielten sehr gute Ergebnisse im Prototyp Grammatiktest "Flurbereinigung" (siehe Tabelle 14). Das am häufigsten erzielte Ergebnis (Modalwert) betrug 100 Prozent, der Mittelwert betrug 88 Prozent, die Hälfte der Teilnehmer erzielte über 93 Prozent (Median). Im Vergleich mit den ausländischen Studienbewerbern werden die Unterschiede besonders deutlich. Die statistischen Kennwerte deuten auf große Unterschiede zwischen den Leistungen deutscher und ausländischer Testteilnehmern (siehe Tabelle 14). Dass diese Unterschiede tatsächlich signifikant sind, kann mit einem *t*-Test für unabhängige Stichproben gezeigt werden (*t*-Test für ungleiche Varianzen: $t(192) = 13{,}719$; $p < 0{,}01$; 2-seitig).

Tabelle 14: Grammatiktest "Flurbereinigung" – Ergebnisse deutscher Studierender und ausländischer Studienbewerber

	deutsche Studierende (n = 76)	**ausländische Studienbewerber (n = 118)**
Mittelwert (arithmetisches Mittel; *AM* in %)	88 %	53 %
Median (Wert, der die Gesamtzahl in zwei Hälften teilt; *Md*)	93 %	53 %
Standardabweichung (Streuungsmaß; s)	13,2 %	21,6 %
Min. – Max.	42 – 100 %	0 – 100 %

Diskussion: Werden im DSH-Grammatiktest Fähigkeiten verlangt, die auch Muttersprachler vor Schwierigkeiten stellen? Die Ergebnisse der Vergleichsuntersuchung mit muttersprachlich deutschen Studierenden sprechen dagegen. Wenn die deutschen Studierenden über eine geringe Punktzahl nicht hinauskamen, waren dafür fast ausnahmslos zwei Ursachen auszumachen: Abbruch des Tests und Rechtschreibfehler. Da der Test am Ende eines Seminars durchgeführt wurde, hörten einige Kandidaten vorzeitig auf und gingen in die Pause. Häufig wurden die letzten Items nicht bearbeitet. Andere Kandidaten erzielten wegen fehlerhafter Rechtschreibung nicht die volle Punktzahl. Typische Fehler waren die Schreibung des s-Lauts (Konjunktion "dass" mit einem "s"; "man lies (!) dabei unbeachtet") oder die Groß- und Kleinschreibung ("veränderungen"). In einigen Fällen scheiterten die deutschen Studierenden an der eigentlichen Aufgabenschwierigkeit des Grammatiktests: "... in der im Zuge der EG Agrarpolitik in den 70er und 80er Jahren vorgenommen (!) Flurbereinigung", "..., der zur (!) nachhaltigen Veränderungen ihrer natürlichen Beschaffenheit führt". Auch wenn Items ausgelassen wurden, könnte dies daran gelegen haben, dass ihnen die geforderte Lösung nicht bekannt war. In den meisten Fällen wurden aber die letzten Items nicht beantwortet. Daher ist anzunehmen, dass der Test abgebrochen wurde. Insgesamt bereiteten das Erkennen der geforderten Transformation und die sprachliche Realisation den deutschen Studierenden kaum Schwierigkeiten.

Bei den Erstsemesterstudierenden technischer Studiengänge ist der Umgang mit Sprache nicht Schwerpunkt des Studiums. Sie waren mit den Aufgaben im DSH-Grammatiktest dennoch nicht überfordert. Der Prototyp-Grammatiktest "Flurbereinigung" stellt besonders fremdsprachige Deutschlerner vor Schwierigkeiten, nicht aber Muttersprachler. Das Testkonstrukt ist ein Indikator für den Grad der Fremdsprachkompetenz.

4.2.3. Ergebnisse der DSH-TestDaF-Pilotstudie

Tabelle 15 gibt einen Überblick über die statistischen Kennzahlen. Im TestDaF-Leseverstehen waren die Ergebnisse am niedrigsten, es folgen die Prüfungsteile zum Hören und zum Schreiben. Im Mündlichen Ausdruck erzielten die Testteilnehmer die höchsten Ergebnisse.

Tabelle 15: DSH-TestDaF-Pilotstudie: Statistische Kennzahlen (n = 56)

Test	Arithmetisches Mittel (*AM*)	Standard-abweichung (s)
TestDaF-HV	3,13	1,06
TestDaF-LV	2,81	0,84
TestDaF-SA	3,48	0,82
TestDaF-MA	3,65	0,99
C-Test	56,6 %	16,8
DSH-Grammatik	48,7 %	23,0 %

Korrelationstests

Die Korrelationen zwischen den Prüfungsteilen der DSH und des TestDaF sind in Tabelle 16 abgebildet. Die Korrelationen wurden mit den Rangkorrelationskoeffizienten von Spearman berechnet.

- **Grammatiktest und C-Test:** Die Ergebnisse im C-Test werden als Indikator für "allgemeine Deutschkompetenz" interpretiert (Begründung siehe Kapitel 6.1.3, Seite 255). Der Zusammenhang zwischen dem Grammatiktest "Flurbereinigung" und dem C-Test liegt im mittleren Bereich (r_s = 0,477; Tabelle 16, Seite 131). Er ist signifikant. Das Streudiagramm (Abbildung 14, Seite 131) verdeutlicht, dass der Zusammenhang zwischen den Ergebnissen der beiden Tests im mittleren Bereich nur sehr begrenzt besteht. Es gibt mehrere Kandidaten, welche im Grammatiktest kaum

Punkte, im C-Test jedoch ein durchschnittliches Ergebnis erzielten. Der Zusammenhang ist auch am Verlauf der Regressionsgeraden zu erkennen: Die Höhenlage der Regressionsgeraden, also der Punkt, an dem die Regressionsgerade die y-Achse schneidet, ist dicht beim Ursprung.

- **Grammatiktest und Schriftlicher Ausdruck:** Der DSH-Grammatiktest weist mittlere Korrelationen mit dem TestDaF Prüfungsteil zum Schreiben auf (Grammatik "Flurbereinigung" und TestDaF-SA: $r_s = 0{,}537$; siehe Tabelle 16, Seite 131).

- **Grammatiktest und Leseverstehen:** Der Grammatiktest korrelierte mittel mit dem TestDaF-Leseverstehen ($r_s = 0{,}411$, siehe Tabelle 16).

- **Grammatiktest und Mündlicher Ausdruck:** Ein signifikanter Zusammenhang zwischen dem Grammatiktest "Flurbereinigung" und dem Prüfungsteil Mündlicher Ausdruck des TestDaF trat nicht auf (siehe Tabelle 16). Alle anderen Korrelations-koeffizienten, die in dieser Phase der Studie berechnet wurden, waren deutlich höher. Das Streudiagramm zeigt, dass die Ergebnisse im mittleren Bereich stark gestreut sind (siehe Abbildung 15, Seite 132). Sowohl Kandidaten mit einem Ergebnis von über 20 Punkten im DSH-Grammatiktest (entspricht mehr als 80 Prozent), als auch Kandidaten mit einem unterdurchschnittlichen Ergebnis im DSH-Grammatiktest erreichten im Mündlichen Ausdruck die TDN 3.

- **Grammatiktest und Hörverstehen:** Die Korrelation zwischen dem Grammatiktest und dem TestDaF-Prüfungsteil zum Hörverstehen ist mittel bis niedrig ($r_s = 0{,}459$; siehe Tabelle 16).

Tabelle 16: DSH-TestDaF-Pilotstudie – Korrelationen

Spearmans rho (r_s)	**C-Test**	**TestDaF LV**	**TestDaF HV**	**TestDaF SA**	**TestDaF MA**
DSH-Grammatik "Flurbereinigung"	**,477**** n=59	**,411**** n=58	**,459**** n=57	**,537**** n=60	**,228** n=55
C-Test		,543** n=62	,589** n=61	,533** n=63	,526** n=57
TestDaF LV			,523** n=61	,440** n=63	,494** n=57
HV				,571** n=61	,718** n=56
SA					,500** n=57
MA					

** Korrelation ist auf dem Niveau von 0,01 signifikant (2-seitig).

* Korrelation ist auf dem Niveau von 0,05 signifikant (2-seitig).

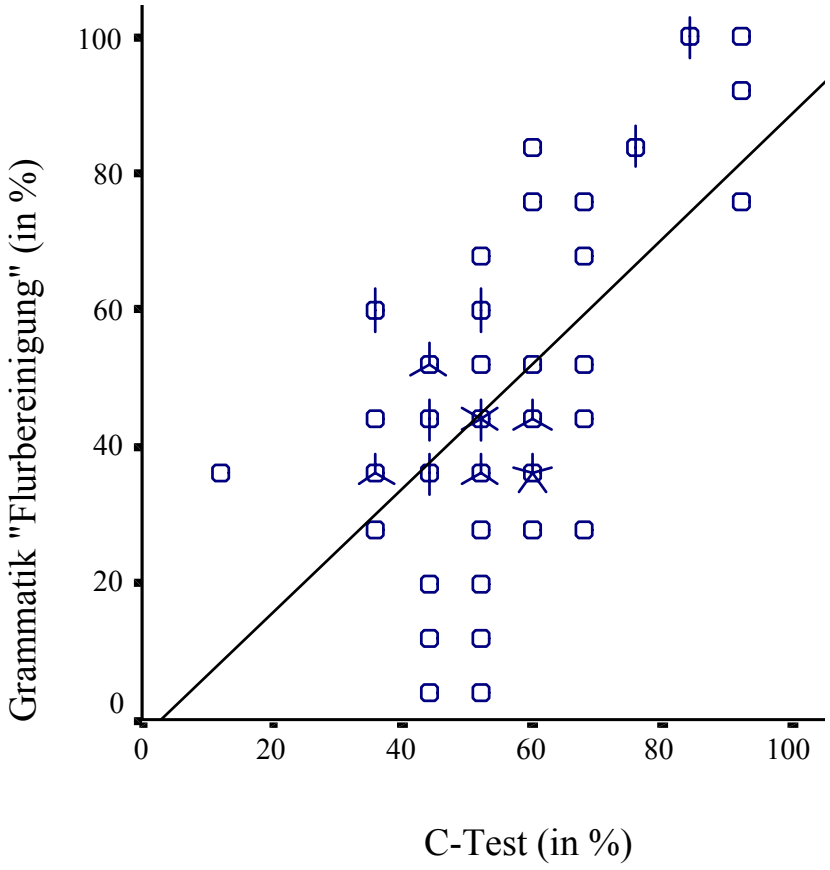


Abbildung 14: Ergebnisse aus dem DSH-Grammatiktest und dem C-Test (Sonnenblumen-Streudiagramm mit linearer Regressionsgeraden; n = 59)

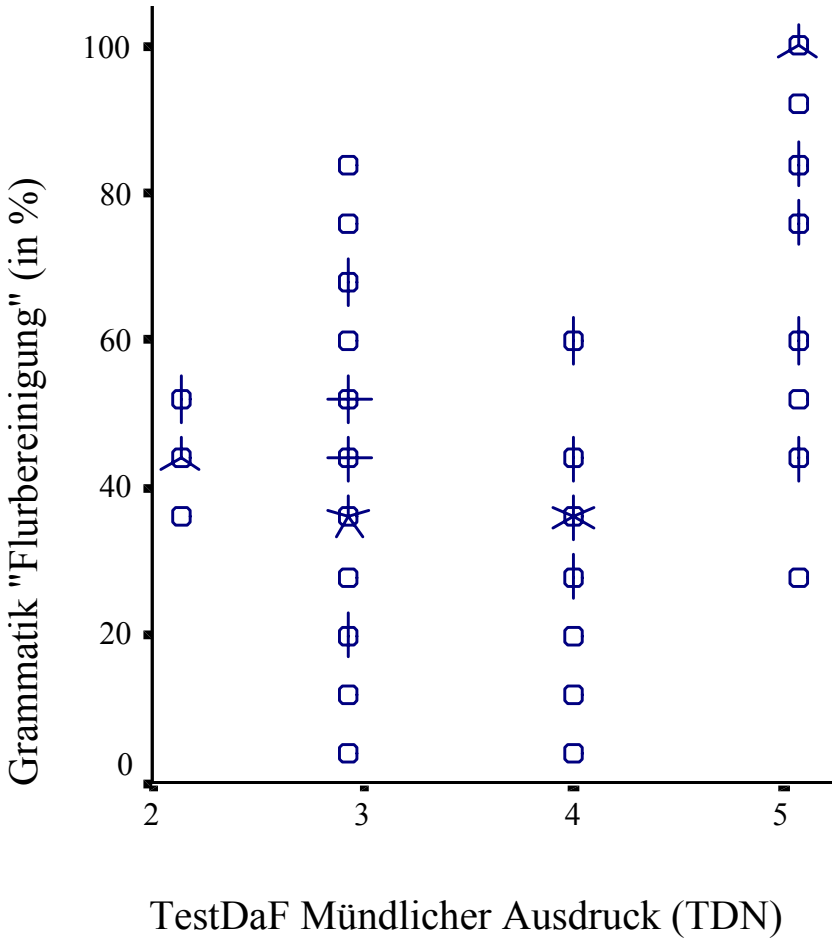


Abbildung 15: Ergebnisse aus dem DSH-Grammatiktest und dem TestDaF-MA (Streudiagramm; n = 55)

Regressionsanalyse

Um die Vorhersagekraft der Leistungen in anderen Prüfungsteilen für den DSH-Grammatiktest zu bestimmen, wurde eine Regressionsanalyse durchgeführt (Backhaus/Erichson/Plinke/Weiber, 2003: 45-116). In die erste Regressionsanalyse wurden die Ergebnisse der Kandidaten im DSH-Grammatiktest "Flurbereinigung" als abhängige Variable (Prognosevariable) und die TestDaF-Prüfungsteile Hörverstehen, Leseverstehen, Schriftlicher Ausdruck und Mündlicher Ausdruck als unabhängige Variablen (Prädiktorvariablen) einbezogen. Auf diese Weise kann untersucht werden, inwieweit die mit den Prüfungsteilen des TestDaF gemessenen Fertigkeiten die Ergebnisse im Grammatiktest "vorhersagen".

Einige Anmerkungen zur Berechnung: Angesichts der relativ kleinen Probandenzahl (n = 55) wurde die Methode "Einschluss" gewählt. Die Variable Grammatiktest wurde mit der Zensur auf einer Skala von 1,0 bis 5,0 eingegeben, die Variablen zum TestDaF mit den TestDaF-Niveaustufen, wobei der Ergebnisklasse "unter TDN 3" die Ziffer "2" zugeordnet wurde. Dabei bin ich von intervallskalierten Daten ausgegangen (siehe Kapitel 4.2.1, Seite 125).

Mit der Methode "Einschluss" konnte ein signifikantes Modell ermittelt werden ($F_{(4,47)} = 11{,}989$; $p < 0{,}01$). Tabelle 17 verdeutlicht, dass die Variable TestDaF-Schriftlicher Ausdruck als signifikanter Prädiktor für die Variable Grammatiktest "Flurbereinigung" interpretiert werden kann. Das gilt auch für die Variable TestDaF-Leseverstehen, allerdings auf einem niedrigeren Signifikanzniveau (95 %). Die TestDaF-Prüfungsteile Hörverstehen und Mündlicher Ausdruck sind keine signifikanten Prädiktoren für den Grammatiktest "Flurbereinigung".

Wie gut kann man das Ergebnis im DSH-Grammatiktest "Flurbereinigung" mithilfe des Modells vorhersagen? An dem Bestimmtheitsmaß, dem korrigierten "R-Quadrat", kann man erkennen, wie gut das Modell die tatsächlichen Ergebnisse abbildet. Das korrigierte R-Quadrat liegt in diesem Fall bei 0,449. Das bedeutet, dass das Modell statistisch 45 Prozent der Varianz der Prognosevariablen erklärt. Es gibt keine allgemein gültige Regel, ab welchem Wert ein R-Quadrat als besonders hoch oder niedrig einzustufen ist. Dies dürfte ein mittlerer Wert sein, der verdeutlicht, dass ein großer Teil der Ergebnisse im Grammatiktest durch das Modell erklärt werden können, aber ein ebenso großer Teil

der Ergebnisse nicht erklärt wird. Im vorliegenden Fall muss man davon ausgehen, dass ein großer Teil der Varianz des DSH-Grammatiktests (über 50 Prozent) durch die anderen Prüfungsteile nicht erfasst wird. Der DSH-Grammatiktest bietet demnach Informationen, welche durch die andere Prüfungsteile nicht erfasst werden.

Tabelle 17: Grammatiktest "Flurbereinigung" – Ergebnisse der Regressionsanalyse (n = 52)

Prädiktor-variablen	nicht-standardisierter Koeffizient B	standardisierter Koeffizient Beta	T	Signifikanz
TestDaF-SA	12,462	0,438	3,136	$p < 0{,}01$
TestDaF-LV	8,961	0,328	2,543	$p = 0{,}014$

Anm.: nicht signifikante Variablen: TestDaF-MA, TestDaF-HV.

Tabelle 18: DSH-TestDaF-Pilotstudie – Faktoranalyse mit TestDaF-Prüfungsteilen und Grammatiktest "Flurbereinigung"

Faktor	Eigenwerte	% der Varianz
1	3,311	66,228
2	,724	14,472
3	,563	11,260
4	,318	6,352
5	,210	4,203

Prüfungsteil	Ladung auf Faktor 1
TestDaF-Hörverstehen	0,901
TestDaF Schriftlicher Ausdruck	0,824
TestDaF Mündlicher Ausdruck	0,803
TestDaF Leseverstehen	0,779
DSH-Grammatik	0,755

Faktoranalyse

Gibt es gemeinsame, hinter den Testergebnissen der einzelnen Prüfungsteile stehende Größen, welche für die Korrelationen verantwortlich sind? Wenn sich die einzelnen Prüfungsteile auf wenige Faktoren reduzieren lassen, könnte die Anzahl der Prüfungsteile möglicherweise ohne Qualitätsverlust reduziert werden und die Struktur der Faktoren gäbe einen Hinweis auf die Konstruktvalidität. Mit den Ergebnissen aus der DSH-TestDaF-Pilotstudie wurde eine Faktoranalyse durchgeführt. Die Eignung der Daten wurde mit Hilfe der Korrelationsmatrix und einer Anti-Image-Korrelations-Matrix festgestellt. Als Extraktionsmethode wurde die Hauptachsen-Faktorenanalyse gewählt, welche von geschätzten Kommunalitäten ausgeht (Backhaus/Erichson/Plinke/Weiber, 2003).

Der erste Faktor erklärt bereits 66 Prozent der Ausgangsvarianz, zusammen mit dem zweiten Faktor erhöht sich der Anteil lediglich um 14 Prozent (siehe Tabelle 18). Die Zahl der zu extrahierenden Faktoren ist nicht festgelegt. Häufig beschränkt man eine Faktoranalyse auf Faktoren, deren Eigenwert über 1 liegt ("Kaiser-Guttman-Kriterium"; Bortz, 1999: 528). Würde man dieses Kriterium anwenden, könnte man nur einen Faktor extrahieren. Auf diesen Faktor "laden" alle fünf Variablen mehr oder weniger hoch, d. h. alle fünf Variablen korrelieren eng mit dem Faktor. Im Falle der DSH-TestDaF-Pilotstudie kann man also nur eine Größe beschreiben, die durch alle Prüfungsteile (TestDaF und DSH-Grammatiktest) bestimmt wird. Es fällt auf, dass der DSH-Grammatiktest am wenigsten auf diesen einen Faktor lädt. Dies sollte jedoch nicht als Hinweis auf die (fehlende) Nützlichkeit des DSH-Grammatiktests interpretiert werden, da die Unterschiede der Faktorladungen nur gering sind.

Dies ist ein typisches Ergebnis, das beispielsweise auch im Rahmen der "*Cambridge-TOEFL Comparability Study*" auftrat. In dieser Studie wurden Faktoranalysen mit den Ergebnissen aus dem TOEFL und dem "*First Certificate in English*" (FCE) durchgeführt. Es ergab sich jeweils nur ein Faktor mit einem Eigenwert über 1 (Bachman/Choi/Davidson/Ryan, 1995: 64-72; Bachman/Davidson/Foulkes/John, 1993). Auch in diesem Fall wurden nur geringe Unterschiede in den Faktorladungen angetroffen. Die Interpretation lautete:

> We feel that at present there is no basis for interpreting this general factor as anything other than a common aspect of language proficiency shared by these subjects as measured by these tests (Bachman/Davidson/Foulkes, 1993: 39).

4.2.4. Ergebnisse der DSH-TestDaF-Vergleichsstudie

Bei der Analyse der Ergebnisse aus der DSH-TestDaF-Vergleichsstudie verfahre ich wie mit den Daten aus der DSH-TestDaF-Pilotstudie. Allerdings liegen hier Daten aus mehr Prüfungsteilen vor: Nicht nur der DSH-Grammatiktest und die Prüfungsteile des TestDaF, sondern auch Ergebnisse in den übrigen DSH-Prüfungsteilen (mit Ausnahme der Mündlichen Prüfung). Zunächst beschreibe ich statistische Kennwerte sowie Korrelationen zwischen den einzelnen Prüfungsteilen.

Wie aus Tabelle 19 hervorgeht, lagen die Ergebnisse im Grammatiktest "Meinungsforschung" unter denen der anderen DSH-Prüfungsteile. Die Streuung der Ergebnisse war breiter als in den übrigen Prüfungsteilen. Der Grammatiktest differenziert stark zwischen den Leistungen der Kandidaten.

Tabelle 19: DSH-TestDaF-Vergleichsstudie – Statistische Kennwerte (n = 56)

	Mittelwert (*AM*)	**Standardabweichung** (s)
DSH-Hörverstehen (Zensur)	3,657	0,99
DSH-Grammatiktest "Meinungsforschung" (Zensur; siehe Abbildung 13)	3,768	1,17
DSH-Leseverstehen (Zensur)	3,459	0,94
DSH-Textproduktion (Zensur)	3,686	1,01
TestDaF LV (TDN)	3,14	0,96
TestDaF HV (TDN)	3,29	0,97
TestDaF SA (TDN)	3,02	0,80
TestDaF MA (TDN)	3,54	0,95

Anm.: DSH: Zensur 1,0 bis 5,0; bestanden bis 4,0. TestDaF: "unter TDN 3" wurde die Ziffer "2" zugeordnet.

Korrelationstests

Betrachtet man die Korrelationen zwischen den Ergebnissen im Grammatiktest "Meinungsforschung" und denen der anderen DSH- und TestDaF-Prüfungsteile, so trifft man auf drei Korrelationen, die auf dem Niveau 99 Prozent signifikant sind: mit den DSH-Prüfungsteilen Leseverstehen und Textproduktion sowie mit dem TestDaF-Prüfungsteil Schriftlicher Ausdruck. Der Rangkorrelationskoeffizient nach Spearman weist mittlere Zusammenhänge des DSH-Grammatiktests "Meinungsforschung" mit den Prüfungsteilen zum Schreiben aus und einen niedrigen Zusammenhang mit dem DSH-Leseverstehen (siehe Tabelle 20). Der niedrige Zusammenhang mit dem DSH-Hörverstehen ist nur auf dem Niveau 95 Prozent signifikant. Signifikante Korrelationen mit den TestDaF-Prüfungsteilen Hörverstehen, Leseverstehen und Mündlicher Ausdruck traten nicht auf.

Tabelle 20: DSH-TestDaF-Vergleichsstudie – Korrelationen (n = 56)

Korrelationen	**DSH HV**	**DSH Gram-matik**	**DSH LV**	**DSH TP**	**TestDaF LV**	**TestDaF HV**	**TestDaF SA**	**TestDaF MA**
Rangkorrelationskoeffizient nach Spearman (obere Dreiecksmatrix)								
DSH HV		**,305***	,361**	,461**	-,409**	-,384**	-,443**	-,321*
DSH Grammatik	**,226***		**,367****	**,474****	**-,134**	**,029**	**-,477****	**-,108**
DSH LV	,270**	**,268****		,327*	-,059	-,161	-,149	-,113
DSH TP	,329**	**,349****	,239**		-,298*	-,242	-,509**	-,253
TestDaF LV	-,329**	**-,103**	-,050	-,241*		,257	,276*	,266*
TestDaF HV	-,314**	**,024**	-,119	-,193	,225*		,350**	,647**
TestDaF SA	-,368**	**-,383****	-,126	-,405**	,236*	,306**		,385**
TestDaF MA	-,251*	**-,078**	-,090	-,208*	,227*	,578**	,338**	
Korrelationen nach Kendall-tau-B (untere Dreiecksmatrix)								

* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.
** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
Anm.: Positive und negative Korrelationen ergeben sich durch unterschiedliche Bewertungssysteme. Eine hohe TDN im TestDaF entspricht einer guten Leistung, eine niedrige Zensur in der DSH verweist auf eine gute Leistung.

Regressionsanalyse

Die 56 Ergebnisse aus der DSH-TestDaF-Vergleichsstudie, welche am Standort Konstanz erhoben wurden, erfüllten die Voraussetzungen für eine Regressionsanalyse. In die Regressionsanalyse wurden die Prüfungsteile aus dem TestDaF und der DSH (mit Ausnahme des Grammatiktests) als Einflussvariablen und der DSH-Grammatiktest als abhängige Variable eingegeben. Mit der Methode "Schrittweise" wurde ein signifikantes Modell ermittelt, welches auf eine nur geringe Vorhersage der Prädiktorvariablen hinweist ($F_{2;53}$ = 11,323; $p < 0{,}01$; Korrigiertes R-Quadrat = 0,273).

Die Regressionsanalyse weist lediglich die Variablen TestDaF-Schriftlicher Ausdruck und DSH-Textproduktion als signifikante Prädiktorvariablen aus (siehe Tabelle 21). Das bedeutet, dass die übrigen Prüfungsteile – also auch die Prüfungsteile zum Leseverstehen – nicht benötigt werden, wenn man die Ergebnisse im DSH-Grammatiktest vorhersagen möchte. Aussagekräftig sind allein die Ergebnisse in den Prüfungsteilen zum Schreiben.

Beachtenswert ist weiter, dass das korrigierte R-Quadrat bei 0,273 liegt. Das bedeutet, dass mit der Regressionsfunktion nur eine Vorhersagewahrscheinlichkeit von ungefähr 27 Prozent erzielt wird. Man würde durch einen Verzicht auf den Grammatiktest nicht alle Informationen verlieren, welche der Grammatiktest für das Ergebnis der Gesamtprüfung liefert, aber doch einen großen Teil. Umgekehrt kann man formulieren, dass sich die Ergebnisse des DSH-Grammatiktests anhand der übrigen Prüfungsteile der DSH und des TestDaF nur zu 27 Prozent vorhersagen lassen.

Die Regressionsfunktion lautet:

DSH-Grammatik = -0,469 * TestDaF-SA + 0,351 * DSH-TP + 3,892.

Tabelle 21: DSH-TestDaF-Vergleichsstudie: Ergebnisse der Regressionsanalyse (n = 56)

Prädiktor-variablen	nicht-standardisierter Koeffizient B	standardisierter Koeffizient Beta	T	Signifikanz
TestDaF-SA	-0,469	-,320	-2,336	$p < 0{,}05$
DSH-TP	0,351	0,303	2,212	$p < 0{,}05$

Faktoranalyse

Wie bereits in der DSH-TestDaF-Pilotstudie führte ich auch mit den Prüfungsteilen der Vergleichsstudie eine Faktoranalyse durch. Mit der Faktoranalyse sollen Zusammenhänge zwischen den Konstrukten einzelner Prüfungsteile entdeckt werden.

Bei der Faktoranalyse wurde die Hauptachsenanalyse angewendet. Aus der Tabelle 22 gehen die Eigenwerte einzelner Prüfungsteile hervor, die hier in Faktoren überführt wurden. Die Eigenwerte stellen die *Gesamt*varianz aller Prüfungsteile dar, die durch einen Faktor/Prüfungsteil aufgeklärt wird. Nach dem Kaiser-Kriterium sollen nur Faktoren mit Eigenwerten extrahiert werden, die größer als 1 sind. Dies gilt in der vorliegenden Faktoranalyse für zwei Faktoren, die zusammen 57 Prozent der Ausgangsvarianz aller einbezogenen Prüfungsteile erklären. Die negative Ladung der TestDaF Prüfungsteile auf diesen Faktor ergibt sich wie bereits bei den Korrelationskoeffizienten durch die unterschiedlichen Bewertungssysteme.

Die varimax-rotierte Faktormatrix weist folgende Faktoren aus: Der erste Faktor ist durch hohe Ladungen der Prüfungsteile DSH-TP, DSH-Grammatiktest und TestDaF-SA sowie niedrige der Prüfungsteile TestDaF-HV und TestDaF-MA gekennzeichnet, der zweite Faktor durch hohe Ladungen der Prüfungsteile TestDaF-HV, TestDaF-MA und DSH-HV sowie niedrige der Prüfungsteile DSH-TP, DSH-LV und DSH-Grammatiktest gekennzeichnet. Erwähnt werden sollten noch die unterschiedlichen Kommunalitäten, zu denen man bei der Extraktion von zwei Faktoren gelangt. Die Kommunalität bezeichnet das Ausmaß, in dem die Varianz *eines* Prüfungsteils durch die beiden Faktoren aufgeklärt wird. Die Varianzanteile der Ergebnisse aus den Prüfungsteilen TestDaF-LV und DSH-LV sind nur zu einem geringen Teil durch die gefundenen Faktoren erklärbar (geschätzte Kommunalitäten auf Basis der Hauptachsenanalyse: TestDaF-LV: 0,169; DSH-LV: 0,208). Die Kommunalitäten der übrigen Prüfungsteile liegen deutlich darüber.

Bei der Faktoranalyse mit den Daten aus der DSH-TestDaF-Vergleichsstudie lassen sich zwei Faktoren ausmachen: Einen Faktor könnte man als Umgang mit geschriebener Sprache identifizieren (Schreiben und Grammatik), den zweiten Faktor als Umgang mit gesprochener Sprache (Hören und Sprechen).

Tabelle 22: DSH-TestDaF-Vergleichsstudie – Faktoranalyse

Extrahierte Faktoren mit Eigenwerten und Varianzerklärungsanteil (Hauptachsen-Faktorenanalyse)				
Faktor	**Eigenwerte**		**% der Varianz**	**Kumulierte %**
1	3,184		39,804	39,804
2	1,401		17,517	57,321
3	,884		11,046	68,366
4	,826		10,319	78,686
5	,556		6,950	85,636
6	,486		6,071	91,707
7	,356		4,453	96,160
8	,307		3,840	100,000
Rotierte Faktorenmatrix bei zwei Faktoren (Kaiser-Kriterium; Varimax-Rotation)				
Prüfungsteil	**Faktor 1**	**Faktor 2**	**Faktor 1 (Rang)**	**Faktor 2 (Rang)**
DSH HV	,501	-,385	4	3
DSH Grammatik "Meinungsforschung"	,696	,035	2	8
DSH LV	,436	-,132	5	7
DSH TP	,738	-,191	1	6
TestDaF LV	-,258	,320	6	4
TestDaF HV	-,061	,911	8	1
TestDaF SA	-,622	,343	3	5
TestDaF MA	-,163	,695	7	2

4.2.5. Zusammenfassung und Diskussion

Fragestellung
Der DSH-Grammatiktest misst produktive Grammatikkompetenz. Welche weiteren Interpretationen des Testergebnisses sind angemessen?

Während die **Korrelationen** des DSH-Grammatiktests mit den Prüfungsteilen Hörverstehen, Leseverstehen und mit dem C-Test auf einem ähnlichen, niedrigen bis mittleren Niveau lagen, lagen die Korrelationen mit dem Schreiben darüber. Nicht signifikant war der Zusammenhang mit dem Sprechen. Dies war sowohl in der DSH-TestDaF-Pilotstudie als auch in der DSH-TestDaF-Vergleichsstudie der Fall. Daraus ist zu schließen, dass sich die Konstrukte der Prüfungsteile zum Sprechen und zur Grammatik deutlich unterscheiden. Dieses Ergebnis steht im Widerspruch zu anderen Studien, in denen hohe Korrelationen von Grammatiktests mit Tests des Leseverstehens beobachtet wurden (z. B. Alderson, 1993).

Die **Regressionsanalysen** mit den Ergebnissen aus der DSH-TestDaF-Pilotstudie und der DSH-TestDaF-Vergleichsstudie verdeutlichten, dass sich die Leistungen im DSH-Grammatiktest statistisch als abhängige Variable von den Leistungen im Schreiben und eventuell auch im Lesen beschreiben lassen. Die Vorhersagewahrscheinlichkeit ist jedoch bei weitem nicht umfassend. Man muss davon ausgehen, dass durch den Grammatiktest ein eigenständiger Beitrag zum Konstrukt eines Sprachtests für den Hochschulzugang geleistet wird, der von den übrigen Prüfungsteilen nur zum Teil erfasst wird, dass ein Grammatiktest also zusätzliche Informationen bietet, die man nicht durch Prüfungsteile zu den sprachlichen Fertigkeiten erschließt.

Die **Faktoranalyse** mit den Daten aus der DSH-TestDaF-Pilotstudie wies lediglich einen Faktor mit einem Eigenwert über 1 aus. Auf diesen Faktor, den man als allgemeine Sprachkompetenz interpretieren könnte, laden alle Tests mehr oder weniger gleich stark, so dass keine weiter gehenden Aussagen gewonnen werden können.

Die Faktoranalyse mit den Ergebnissen aus der DSH-TestDaF-Vergleichsstudie ermöglicht folgende Interpretation: Unter Berücksichtigung des Kaiser-Kriteriums (Eigenwerte größer als 1) wurden zwei Faktoren extrahiert. Auf den ersten Faktor laden neben dem Grammatiktest vor allem die produktiven Prüfungsteile DSH-Textproduktion sowie TestDaF-Schriftlicher Ausdruck. Dies deutet darauf hin, dass die Fähigkeiten, die man in den im Grammatiktest geforderten Transformationen einsetzen muss, auch beim Schreiben eingesetzt werden. Weniger haben sie zu tun mit dem zweiten Faktor, der durch hohe Ladungen des Hörverstehens und des Mündlichen Ausdrucks gekennzeichnet ist.

Bei einer Interpretation der Ergebnisse aus den statistischen Analysen ist zu berücksichtigen, dass sie jeweils auf Korrelationskoeffizienten beruhen. Mit Korrelationen kann man zwar die Stärke eines Zusammenhangs aufdecken, nicht jedoch eine Kausalität. Ein ursächlicher Zusammenhang, wie er für die Konstruktdefinition nötig wäre, lässt sich daher nicht allein mit den verwendeten statistischen Verfahren begründen, sondern nur zusammen mit inhaltlichen Überlegungen (siehe auch Kapitel 2.2, Seite 49 ff). Die Ergebnisse der Studie lassen vor diesem Hintergrund folgende Interpretationen als gesichert erscheinen:

- Der DSH-Grammatiktest bietet Informationen, die nicht durch Tests zu den sprachlichen Fertigkeiten erfasst werden. Darauf deuten die Regressionsanalysen hin.

- Fähigkeiten im Umgang mit gesprochener Sprache (Hören, Sprechen) werden vom DSH-Grammatiktest nicht oder nur in geringem Maße erfasst. Der DSH-Grammatiktest ist daher kein geeigneter Test, um "allgemeine Sprachfähigkeit" zu erfassen.

Es konnte ein starker Zusammenhang zwischen dem DSH-Grammatiktest und Prüfungsteilen zum Schreiben festgestellt werden. Dies könnte als Hinweis auf eine Ähnlichkeit der Testkonstrukte interpretiert werden. Es ist denkbar, dass bei der Aufgabe, Sätze bzw. einen Text zu produzieren, vergleichbare Fertigkeiten geprüft werden wie bei der Aufgabe, kontextualisierte Sätze zu transformieren. Hier spiegeln sich auch die Leitlinien für die Bewertung des Schriftlichen Ausdrucks, bei der die sprachliche Korrektheit ein wichtiges Kriterium darstellt.

Wie legitim ist der DSH-Grammatiktest als Teil eines Sprachtests für den Hochschulzugang? Der DSH-Grammatiktest bietet durch das Testkonstrukt "produktive Grammatikkompetenz" Informationen, die nicht durch Prüfungsteile zu den sprachlichen Fertigkeiten erfasst werden. Ob diese Informationen für den Testanwender von Nutzen sind, wurde bislang nicht beantwortet. Offen ist ebenso, ob die Fähigkeit "produktive Grammatikkompetenz" für ausländische Studierenden eine Hilfe darstellt, um im Fachstudium bestehen zu können.

4.3. Auswirkungen auf die Zulassungs-entscheidung

Übersicht: Kapitel 4.3
In diesem Kapitel gehe ich der Frage nach, welche Auswirkungen der DSH-Grammatiktest auf die Zulassungsentscheidung hat. Es wird untersucht, ob der DSH-Grammatiktest für die Testteilnehmer ein "Joker" ist. Ich gehe am Rande auch auf die Frage ein, ob sich die Überlegungen zur Konstruktvalidität (Kapitel 4.2) nachvollziehen lassen. Ich greife dazu auf die Ergebnisse der DSH-TestDaF-Pilotstudie sowie auf Ergebnisse der DSH-TestDaF-Vergleichsstudie zurück.

4.3.1. Fragestellung und Methode

In diesem Abschnitt geht es um den praktischen Nutzen des Grammatiktests für die Zulassungsentscheidung. In Kapitel 2.1 (Seite 13 ff) argumentierte ich, dass die Auswahl von Kandidaten aus einer größeren Gruppe eine zentrale Funktion der DSH darstellt. Die Zulassungsfunktion ist darüber hinaus durchaus von den Testerstellern beabsichtigt.

Ein DSH-Grammatiktest bietet also zusätzliche Informationen, welche nicht durch andere Prüfungsteile erfasst werden. Benötigt man die zusätzlichen Informationen aus dem DSH-Grammatiktest für die Entscheidung über den Hochschulzugang? Stellen sie eine notwendige oder überflüssige Information dar? Zu klären ist, ob sich bei einem Verzicht auf den Grammatiktest ein Verlust an Informationen ergeben würde oder ob die Informationen aus den übrigen Prüfungsteilen ausreichen, um sich ein Bild über den Sprachstand der Kandidaten zu machen. Die Nützlichkeit der Prüfung wäre erhöht, wenn durch den DSH-Grammatiktest relevante Informationen über den Sprachstand der Kandidaten liefert, die man nicht aus den anderen Prüfungsteilen entnehmen kann.

Fragestellungen

Benötigt man die Informationen aus dem DSH-Grammatiktest für die Zulassungsentscheidung? Welche Auswirkungen hat der DSH-Grammatiktest auf die Zulassungsentscheidung?

Der Fragenkomplex weist eine Nähe zum Thema der Konstruktvalidität auf. Dort wurden statistische Verfahren eingesetzt (siehe Kapitel 4.2). In diesem Teil der Untersuchung wird eine qualitativ ausgerichtete Analyse auffälliger Ergebnisse aus der DSH-TestDaF-Pilotstudie und aus der DSH-TestDaF-Vergleichsstudie durchgeführt. Mit einer Kombination von quantitativen und qualitativen Methoden soll ein umfassenderes Bild entstehen als dies allein mit statistischen Untersuchungen an den Stichproben möglich wäre. Die Auswirkungen auf die Zulassungsentscheidung werden anhand der Ergebnisse einzelner Kandidaten erläutert.

An den Ergebnissen einzelner Kandidaten wird die Auswirkung des Grammatiktests auf die Gesamtnote beschrieben. Wenn das Ergebnis des Grammatiktests von denen der anderen DSH-Prüfungsteile abweicht, ist der Einfluss besonders groß. Bei einem überdurchschnittlichen Ergebnis verbessert sich das Gesamtergebnis, bei einem unterdurchschnittlichen verschlechtert es sich. Daher analysiere ich die Ergebnisse von Kandidaten mit "auffälligen" Ergebnissen im Grammatiktest genauer und frage, ob der Einfluss des Grammatiktests auf die Zulassungsentscheidung legitim und wünschenswert ist. Die DSH-TestDaF-Pilotstudie und die DSH-TestDaF-Vergleichsstudie wurden bereits in Kapitel 4.2.1 (Seite 116 ff) beschrieben. Mit einzelnen Kandidaten aus der DSH-TestDaF-Pilotstudie wurden kurze Gespräche über ihre Lernbiographie geführt. Die Kandidaten aus der DSH-TestDaF-Vergleichsstudie füllten außerdem einen umfangreichen Fragebogen zu ihrer Lernbiographie aus (Abdruck in: Koreik, 2005).

4.3.2. Ergebnisse und Diskussion

Zunächst stelle ich Ergebnisse von Kandidaten aus der DSH-TestDaF-Pilotstudie dar, anschließend gehe ich auf die Ergebnisse der DSH-TestDaF-Vergleichsstudie ein.

DSH-TestDaF-Pilotstudie

Die Studie befasste sich mit Kandidaten, deren Ergebnisse im Grammatiktest auffällig waren (siehe Tabelle 23, Seite 149). Bei Kandidaten mit einem Ergebnis im Grammatiktest, welches unter dem anderer Prüfungsteile lag, konnten zwei Gruppen unterschieden werden:

- **Sprechgewandte 1:** Die "Sprechgewandten 1" hatten nicht allein im Grammatiktest, sondern auch im Umgang mit geschriebener Sprache Schwierigkeiten.
- **Sprechgewandte 2:** Bei den "Sprechgewandten 2" gaben die überdurchschnittlichen Ergebnisse in allen Prüfungsteilen des TestDaF keine Hinweise auf sprachliche Defizite. Allein das Ergebnis im Grammatiktest war unterdurchschnittlich. Bei dieser Gruppe kann die Varianz nicht durch die übrigen Prüfungsteile erklärt werden, der Grammatiktest bietet zusätzliche Informationen.
- **Grammatikfreaks:** Außerdem fallen Ergebnisse von Kandidaten mit hohen Leistungen im Grammatiktest auf, die in anderen Prüfungsteilen nicht wiederholt werden. Kandidaten mit einem guten Ergebnis im Grammatiktest, aber durchschnittlichen oder schwachen Ergebnissen in den anderen Prüfungsteilen, bezeichne ich als "Grammatikfreaks".

Die "**Sprechgewandten 1**" haben Schwierigkeiten im Grammatiktest und im Umgang mit geschriebener Sprache. Bei dieser Gruppe von Kandidaten fällt das niedrige Ergebnis im DSH-Grammatiktest auf, niedrig sind auch die Ergebnisse in Prüfungsteilen zum Schreiben und zum Lesen.

Die Kollegiatin aus der Slowakei (1) erzielte ein durchschnittliches Ergebnis im Prototyp-Grammatiktest "Flurbereinigung", im TestDaF MA und im HV erreichte sie jedoch die höchste Niveaustufe und im LV "unter TDN 3". Etwas abweichend von anderen "Sprechgewandten" erzielte sie im Prüfungsteil TestDaF SA die TDN 4. Wie sind diese uneinheitlichen Ergebnisse zu erklären? Die Kollegiatin verfügt bereits über eine abgeschlossene Berufsausbildung aus ihrem Heimatland. Deutsch lernte sie in der Schule und auch über deutsches Fernsehen, das sie als Jugendliche regelmäßig schaute. Die Ergebnisse in den Grammatiktests überraschten sie nicht, denn sie hatte "immer schon Schwierigkeiten mit Grammatik". Im ersten Semester des Studienkollegs hatte sie in den nicht-sprachlichen Fächern Schwierigkeiten, wurde aber in das zweite Semester versetzt. Benötigt man den DSH-Grammatiktest, um auf die Tatsache aufmerksam zu werden, dass sie zwar fließend deutsch sprechen kann, bei anderen Fertigkeiten ein vergleichbar hohes Niveau jedoch nicht immer erreicht? Meiner Ansicht nach gibt der TestDaF genügend Hinweise darauf, denn im TestDaF LV verfehlte sie die TDN 3 mit nur acht von 30 möglichen Punkten deutlich (TDN 3 ab 15 Punkten). Mit dem TestDaF-Zeugnis wäre sie nicht zu einem Studium zugelassen worden.

Mehrere Kandidaten erzielten im DSH-Grammatiktest ein unterdurchschnittliches Ergebnis (fast keine richtige Lösung), im Mündlichen Ausdruck des TestDaF jedoch die TDN 4. Dazu gehört ein Kandidat aus dem Libanon, der sich zum Testzeitpunkt am Beginn des zweiten Semesters des Studienkollegs befand (Technikkurs).

Der 22-jährige Libanese ist vielsprachig aufgewachsen. In Venezuela wurde er geboren und er hat bereits in mehreren Ländern gelebt. Er beherrscht neben seiner Muttersprache Arabisch auch Englisch, Französisch, Spanisch und natürlich Deutsch. Grundstufenkurse für den Deutscherwerb belegte er im Libanon, obwohl sein Deutscherwerb seiner Ansicht nach eigentlich erst in Deutschland begonnen habe. Die guten Erfahrungen, die er beim Fremdsprachenerwerb in fremdsprachlicher Umgebung gemacht hat, haben zu einer bewussten Vernachlässigung des strukturierten Spracherwerbs geführt. Dies manifestiert sich in den unterdurchschnittlichen Leistungen in den Prüfungsteilen TestDaF-LV (TDN 3), TestDaF SA ("unter TDN 3") und vor allem im Grammatiktest (11 %). Seine guten Leistungen im TestDaF Hörverstehen (TDN 4) und im Mündlichen Ausdruck (TDN 4) geben Hinweise auf seine fortgeschrittenen Fähigkeiten im Umgang mit gesprochener Sprache.

Tabelle 23: DSH-TestDaF-Pilotstudie – auffällige Ergebnisse im Grammatiktest

	TestDaF LV, HV, SA, MA (TDN)	**C-Test (%)**	**DSH-Grammatik "Flurbereinigung"**
AM (n = 56)		57 %	49 %
Die "Sprechgewandten 1": Schwierigkeiten im Grammatiktest und im Umgang mit geschriebener Sprache			
Kollegiatin 1 (Slowakei)	unter 3-5-4-5	69 %	47 %
Kollegiat (Libanon)	3-4-unter 3-4	53%	11%
Die "Sprechgewandten 2": Schwierigkeiten im Grammatiktest, aber kein Hinweis auf Defizite im TestDaF			
Kollegiatin (Peru)	4-5-4-5	69%	30%
Kollegiatin 2 (Slowakei)	4-5-5-5	63%	42%
Die "Grammatikfreaks": im Sprechen und Hören eher unterdurchschnittlich			
Kollegiat (China)	4-3-3-3	63%	81%

Anmerkung: TestDaF-Ergebnisse in der Reihenfolge LV, HV, MA, SA.

Bei der Gruppe von Kandidaten, die einen unterdurchschnittlichen DSH-Grammatiktest, aber ein überdurchschnittliches Ergebnis im Mündlichen Ausdruck erzielten, lässt sich ein Muster erkennen: Sie halten sich bereits einige Zeit in Deutschland auf, der Anteil des ungesteuerten Spracherwerbs ist hoch. Wenn die schlechten Leistungen im DSH-Grammatiktest ein Indikator für unzureichende sprachliche Leistungen sind, gibt es mindestens einen anderen Prüfungsteil des TestDaF, welcher ebenfalls auf Defizite hinweist. Die Ergebnisse der "Sprechgewandten 1" bestätigen die Annahmen zum Konstrukt des DSH-Grammatiktests: Schwierigkeiten im Umgang mit geschriebener Sprache (Lesen/Schreiben) spiegeln sich in unterdurchschnittlichen Leistungen im DSH-Grammatiktest. Der Umgang mit gesprochener Sprache entwickelte sich bei dieser Gruppe relativ unabhängig von den Grammatikkenntnissen.

Die "**Sprechgewandten 2**", die nächste Gruppe, die ich betrachten möchte, haben Schwierigkeiten im Grammatiktest, es gibt aber im TestDaF keinen Hinweis auf Defizite. Die Kandidaten erzielten ein niedriges Ergebnis im Grammatiktest "Flurbereinigung", aber ein hohes Ergebnis in allen TestDaF-Prüfungsteilen, vor allem im Mündlichen Ausdruck. Hierzu zählen eine Kollegiatin aus Peru und eine aus der Slowakei (2).

Die Peruanerin erzielte zweimal die TDN 4 und zweimal die TDN 5. Nur im Grammatiktest ist ihr Ergebnis unterdurchschnittlich. Liegt hier ein Hinweis auf eine mangelnde Studierfähigkeit vor, die durch den TestDaF nicht aufgedeckt wird? Ihre Sprachlernbiografie und die übrigen Leistungen im Studienkolleg deuten nicht darauf hin. In den zwei Jahren vor ihrer Ausreise nach Deutschland besuchte sie eine zweisprachige Schule. Ebenfalls in Peru erhielt sie das Kleine Deutsche Sprachdiplom (KDS). Sie absolvierte den TestDaF zu Beginn des ersten Semesters des Studienkollegs. Zu dem Zeitpunkt konnte sie bereits flüssig und verständlich sprechen. Auch in den anderen TestDaF-Prüfungsteilen und im C-Test stellte sie ihre fortgeschrittenen Deutschkenntnisse unter Beweis. Mit Ausnahme des DSH-Grammatiktests lagen ihre Ergebnisse in den übrigen Testteilen durchweg deutlich über dem Durchschnitt. Im Gespräch gab sie an, dass sie das Testformat des DSH-Grammatiktests verwirrt habe und sie nicht gewusst habe, was von ihr verlangt wird. Die sprachlichen Phänomene, die für die geforderten Transformationen notwendig sind, seien ihr eigentlich nicht fremd gewesen, was sie bei der Besprechung des Tests bemerkt habe. In ihrem Fall liefert ein unvorbereiteter DSH-Grammatiktest zwar die zusätzliche Information, dass sie Schwierigkeiten mit dem Testformat hat. Schlussfolgerungen für die Studierfähigkeit lassen sich daraus nicht ziehen, denn ihrer Interpretation des Testergebnisses ist wahrscheinlich zuzustimmen. Sie müsste sich mit dem Testformat vertraut machen, dann würde sie in diesem Testteil ein besseres Ergebnis erzielen.

Bei der Kollegiatin aus der Slowakei (2) ist das Prüfungsergebnis ähnlich. Die 26-jährige Kollegiatin lebte vor dem Besuch des Studienkollegs bereits zwei Jahre als Au-pair-Mädchen in Deutschland. Bei diesem hohen Anteil an ungesteuertem Spracherwerb überrascht es nicht, dass die mündliche Sprachkompetenz besser ausgeprägt ist als die Fähigkeit, den DSH-Grammatiktest erfolgreich zu absolvieren.

Im Fall der beiden Kandidatinnen, die zu der Gruppe der "Sprechgewandten 2" gezählt wurden, wäre eine Interpretation des unterdurchschnittlichen Ergebnisses im Grammatiktest als Hinweis auf Defizite der Sprachbeherrschung in Bezug auf die Studierfähigkeit meiner Ansicht nach unzutreffend.

Die "**Grammatikfreaks**" haben im Sprechen und Hören eher unterdurchschnittliche Ergebnisse. Sie erzielten ein hohes Ergebnis im Grammatiktest, aber ein niedriges im Mündlichen Ausdruck. Als Beispiel sei der Kollegiat aus China angeführt. Der 21-

jährige Chinese studierte in China zwei Semester Chemie – was ihn allerdings nicht interessierte. In der Schule und während des Studiums lernte er Englisch, aber noch kein Deutsch. Seinen Wunsch, ein Studium in Großbritannien oder in den USA, konnte er aus finanziellen Gründen nicht verwirklichen. Ein Jahr vor der Teilnahme an der Prüfung kam er ohne Deutschkenntnisse nach Deutschland und nahm an einem einjährigen Intensivkurs an einer privaten Sprachschule teil. Der Intensivkurs schloss mit einer Mittelstufenprüfung ab. Wegen seiner sehr guten Leistungen im (schriftlichen) Aufnahmetest zum Studienkolleg wurde er gleich in das zweite Semester des Studienkollegs aufgenommen. Der junge Mann kann auf einen strukturierten Deutscherwerb in Deutschland zurückblicken. Dies scheint ihm allerdings nicht beim Mündlichen Ausdruck und auch nicht beim Hörverstehen zu helfen. Der Kollegiat verfügt offensichtlich über besondere Fertigkeiten, die er im DSH-Grammatiktest unter Beweis stellen kann. Sie manifestieren sich auch beim Lesen, nicht jedoch beim Sprechen und Hören, in diesem Fall auch nicht beim Schreiben.

DSH-TestDaF-Vergleichsstudie

Auch im Umgang mit den Daten aus der DSH-TestDaF-Vergleichsstudie (siehe Kapitel 4.2.1, Seite 118 ff) wähle ich Ergebnisse von Kandidaten aus, welche ein auffälliges Ergebnis im Grammatiktest haben. Eine Anmerkung dazu: Würde man das Ergebnis der Schriftlichen Prüfung der DSH ohne die Ergebnisse aus dem Grammatiktest bilden, so hätte das für die Mehrheit der 56 Kandidaten eine positive Auswirkung. Dies liegt an dem im Mittel niedrigen Ergebnis im Grammatiktest im Vergleich mit den anderen Prüfungsteilen.

In der Tabelle 24 (Seite 154) sind die Ergebnisse derjenigen Kandidaten abgebildet, bei denen sich das Ergebnis des Grammatiktests besonders negativ oder besonders positiv auf das Endergebnis auswirkt. Die Spalte "mit Grammatik" zeigt die Zensur der Schriftlichen Prüfung unter Einbeziehung des Grammatiktests. Die Spalte "ohne Grammatik" enthält ein fiktives Ergebnis der Schriftlichen Prüfung, das nur aus dem Mittelwert der Prüfungsteile Hörverstehen, Leseverstehen und Textproduktion gebildet wurde – also ohne Berücksichtigung des Grammatiktests und bei gleicher Gewichtung der Prüfungsteile. Aus der Spalte "Differenz" geht der Einfluss des Grammatiktests auf

das Ergebnis hervor. Da Schriftliche und Mündliche Prüfungen unabhängig voneinander gewertet werden, fließen die Ergebnisse der Mündlichen Prüfungen nicht mit ein.

Zunächst betrachte ich die Ergebnisse der Kandidaten, bei denen der Grammatiktest einen negativen Einfluss auf das Endergebnis ausübt. Aus der Gruppe der "**Sprechgewandten 1**" stelle ich einige Kandidatinnen und Kandidaten vor.

Die Kandidatin 24 ist eine 24-jährige Studienbewerberin aus Rumänien mit dem Studienziel "Informatik". Sie kann zur Gruppe der "Sprechgewandten 1" gezählt werden. Deutsch ist ihre dritte Fremdsprache. Mit dem Deutscherwerb begann sie eineinhalb Jahre vor dem Testzeitpunkt in Deutschland. Es gibt einen deutlichen Unterschied zwischen den geringen Leistungen im Grammatiktest und den Leistungen in den übrigen Prüfungsteilen der DSH bzw. des TestDaF. Von Interesse ist weiter, dass sie im TestDaF Leseverstehen und im Schriftlichen Ausdruck nur die TDN 3, im Hörverstehen und Mündlichen Ausdruck aber die TDN 4 bzw. 5 erzielte. Dies spiegelt die Erkenntnisse der Faktoranalysen: Ihre Stärken liegen laut TestDaF im Faktor, der von der gesprochenen Sprache bestimmt wird. Ihre Schwächen liegen in der Schriftsprache und der Grammatik. Dies wird in der DSH durch das niedrige Ergebnis im Grammatiktest deutlich. Aus den übrigen Ergebnissen aus der DSH hätte man diese Informationen nicht erschließen können. Von Interesse ist weiter, dass sie sich laut Selbstauskunft besonders auf den Grammatiktest vorbereitet hat. Offensichtlich war sie sich über ihre Schwäche bewusst, konnte sie jedoch nicht beheben.

Das Muster der "Sprechgewandten 1" zeigt sich auch bei der Kandidatin 13, eine in Deutschland verheiratete Indonesierin, oder des Kandidaten 42, einem 23-jährigen Tunesier, zu. Bei diesen Kandidaten spiegeln sich die Annahmen zur Nähe des Konstrukts des DSH-Grammatiktests zum Schreiben . Mit Blick auf das Lesen bestätigt die Betrachtung einzelner Ergebnisse die in statistischen Analysen gefundene Beobachtung, dass die Konstrukte des Lesens und des Grammatiktests nicht äquivalent sind. Die Kandidatin 47 erzielte beispielsweise im Grammatiktest nur ein niedriges Ergebnis, in den DSH- und TestDaF-Prüfungsteilen zum Lesen aber überdurchschnittliche Ergebnisse.

Die folgenden Kandidaten sind eher den "**Sprechgewandten 2**" zuzurechnen (niedriges Ergebnis im Grammatiktest, aber keine Hinweise auf Defizite in anderen Prüfungsteilen). Bei der Kandidatin 29 liegt das Ergebnis im Grammatiktest deutlich unter den Ergebnissen aus den anderen DSH-Prüfungsteilen. Im TestDaF lagen die Ergebnisse im mittleren Bereich (TDN 3-4-4-4). Die 19-jährige Studienbewerberin aus Estland lernte bereits seit drei Jahren Deutsch und nahm in Deutschland an Kursen des Goethe-Instituts teil. Ihrer eigenen Einschätzung nach ist ihre Leistung im Lesen besser als in den übrigen Fertigkeiten, was sich allerdings in den Tests nicht spiegelt..

Der Kandidat 5 ist ein 29-jähriger Kameruner, der an Goethe-Instituten im Heimatland und in Deutschland Deutsch lernte. Er schätzt seine Grammatikkenntnisse als sehr gut ein, erzielt jedoch nur ein niedriges Ergebnis. Seine Ergebnisse in den übrigen DSH- und TestDaF-Prüfungsteilen liegen im mittleren Bereich.

Bei den "Sprechgewandten 2" (d. h. die Kandidaten, deren Ergebnisse im oberen Teil der Tabelle 24 abgebildet sind) liefert der Grammatiktest zusätzliche Informationen, die sich nur zum Teil aus den Ergebnissen der übrigen Prüfungsteile ableiten lassen.

Nun betrachte ich die Ergebnisse der "**Grammatikfreaks**", bei denen der Grammatiktest einen besonders positiven Einfluss auf das Gesamtergebnis ausübt. Im Fragebogen gaben zehn der zwölf Kandidaten an, dass sie sich besonders intensiv auf den Grammatiktest vorbereitet hätten. Bei ihnen hat sich die Vorbereitung ausgezahlt. Da der Grammatiktest schwieriger war als die übrigen Prüfungsteile, ist der Effekt nicht so ausgeprägt wie im umgekehrten Fall.

Der Kandidat 8 profitiert am stärksten vom Grammatiktest. Der 20-jährige Tunesier, der ein Architekturstudium aufnehmen möchte, erzielt nur im Grammatiktest ein deutlich überdurchschnittliches Ergebnis, in den übrigen Prüfungsteilen sind seine Ergebnisse durchschnittlich bis unterdurchschnittlich. Nur im TestDaF-Prüfungsteil Mündlicher Ausdruck erzielt er mit der TDN 4 noch ein überdurchschnittliches Ergebnis. Er lernte drei Jahre Deutsch in der Schule in Tunesien und besuchte insgesamt 14 Monate einen Sprachkurs in Deutschland.

Tabelle 24: DSH-TestDaF-Vergleichsstudie – auffällige Ergebnisse im Grammatiktest

Nr.	DSH HV	DSH Grammatik	DSH LV	DSH TP	DSH *AM* ohne Grammatik	DSH *AM* mit Grammatik	Differenz	TestDaF LV	TestDaF HV	TestDaF SA	TestDaF MA	TestDaF Mittelwert
negative Beeinflussung des Gesamtergebnis durch den Grammatiktest: die "**Sprechgewandten**"												
24	3,0	**4,8**	2,1	2,8	2,6	3,2	-0,6	TDN 3	TDN 4	TDN 3	TDN 5	3,75
29	1,7	**4,9**	3,8	3,5	3,0	3,5	-0,5	TDN 3	TDN 4	TDN 4	TDN 4	3,75
5	3,1	**4,8**	3,8	2,2	3,0	3,5	-0,5	TDN 4	TDN 3	TDN 4	TDN 4	3,75
10	3,2	**5,0**	3,8	3,0	3,3	3,8	-0,5	TDN 5	TDN 4	TDN 4	TDN 4	4,25
7	2,4	**5,0**	4,0	4,0	3,5	3,9	-0,4	TDN 4	TDN 3	unter 3	TDN 3	3,00
42	3,0	**4,8**	3,4	3,6	3,3	3,7	-0,4	TDN 4	TDN 4	unter3	TDN 4	3,50
47	4,1	**4,6**	2,6	3,0	3,2	3,6	-0,4	TDN 5	TDN 3	TDN 3	TDN 4	3,75
13	3,2	**4,9**	3,6	4,1	3,6	4,0	-0,4	TDN 4	TDN 4	TDN 3	TDN 3	3,50
27	3,0	**5,0**	3,6	4,4	3,7	4,0	-0,3	unter3	TDN 3	TDN 3	TDN 3	2,75
35	3,9	**4,5**	2,4	3,5	3,3	3,6	-0,3	TDN 3	TDN 3	unter3	TDN 3	2,75
15	3,9	**5,0**	2,7	4,8	3,8	4,1	-0,3	unter3	unter3	TDN 3	TDN 3	2,50
positive Beeinflussung des Gesamtergebnisses durch den Grammatiktest: die "**Grammatikfreaks**"												
8	3,3	**1,5**	5,0	3,8	4,0	3,4	+0,6	TDN 3	TDN 3	TDN 3	TDN 4	3,25
17	4,7	**1,8**	2,7	5,0	4,1	3,6	+0,5	TDN 3	TDN 3	TDN 3	TDN 5	3,50
14	3,2	**1,6**	3,8	2,5	3,2	2,8	+0,4	TDN 3	unter3	TDN 4	TDN 4	3,25
40	2,2	**1,8**	3,8	3,7	3,2	2,9	+0,3	TDN 4	TDN 4	TDN 4	TDN 4	4,00
22	4,3	**2,6**	3,7	3,8	3,9	3,6	+0,3	unter3	TDN 3	TDN 3	TDN 3	2,75
43	4,0	**1,7**	2,8	2,2	3,0	2,7	+0,3	TDN 4	TDN 3	TDN 4	TDN 4	3,75
48	5,0	**3,7**	5,0	5,0	5,0	4,7	+0,3	TDN 5	unter3	TDN 3	TDN 3	3,25
50	3,7	**2,0**	2,0	3,2	3,0	2,7	+0,3	unter3	TDN 4	TDN 3	TDN 3	3,00
31	3,6	**3,0**	3,5	4,5	3,9	3,7	+0,2	TDN 3	TDN 3	TDN 3	TDN 5	3,50
11	3,5	**2,3**	3,6	2,2	3,1	2,9	+0,2	TDN 3	TDN 3	TDN 3	TDN 3	3,00
37	4,5	**4,0**	4,8	5,0	4,8	4,6	+0,2	unter3	unter3	TDN 3	unter3	2,25
12	3,9	**2,7**	3,1	3,0	3,3	3,2	+0,1	TDN 4	TDN 3	TDN 4	TDN 3	3,50

Anm.: TestDaF-Ergebnisklassen als TDN (TestDaF-Niveaustufen); "unter TDN 3" wurde "unter 3" zugeordnet. Für die Berechnung des Mittelwerts wurde "unter TDN 3" als "2" codiert.
DSH-Ergebnisklassen als Zensur (1,0 – 5,0, bestanden bis 4,0)

Der 22-jährige Kameruner (Kandidat 17) strebt ein technisches Studium an. Durch das gute Ergebnis im DSH-Grammatiktest bestand er die Schriftlichen Prüfung der DSH. Ohne den Grammatiktest wäre er durchgefallen. Er lernte zweieinhalb Jahre in seinem Heimatland Deutsch und insgesamt sieben Monate an Sprachschulen in Deutschland.

Bei den Kandidaten 14, 22, 50, 31 und 11 handelt es sich um Chinesinnen und Chinesen, die Kandidaten 48 und 37 stammen aus Vietnam. Diese Kandidaten mit entfernten Muttersprachen können den DSH-Grammatiktest nutzen, um bestimmte Schwächen auszugleichen. Im TestDaF-Prüfungsteil Hörverstehen erzielten sie beispielsweise mit einer Ausnahme nur die TDN 3 oder "unter TDN 3". Dennoch reichen die Leistungen im DSH-Grammatiktest nicht aus, um das Ergebnis im Schriftlichen Teil der DSH so weit zu verändern, dass die Zulassungsentscheidung beeinflusst wird.

Die Ergebnisse der "Grammatikfreaks" aus der DSH-TestDaF-Vergleichsstudie sind uneinheitlich. Die Leistungen im Grammatiktest liegen über den Leistungen in anderen Prüfungsteilen. Nur bei wenigen Kandidaten sind auch die Ergebnisse im Schreiben (oder im Lesen) überdurchschnittlich.

4.3.3. Zusammenfassung und Diskussion

Fragestellungen
Benötigt man die Informationen aus dem DSH-Grammatiktest für die Zulassungsentscheidung? Welche Auswirkungen hat der DSH-Grammatiktest auf die Zulassungsentscheidung?

Ist es für die Zulassungsentscheidung notwendig, über schlechte Leistungen einiger Kandidaten ("Sprechgewandte") im DSH-Grammatiktest zu erfahren, wenn die Ergebnisse in den anderen Prüfungsteilen keine oder nur wenige Hinweise auf Defizite gaben? Bei einigen Probanden aus den beiden Studien trat dieses Phänomen auf:

Bei der Gruppe der "Sprechgewandten 1" waren die Ergebnisse in den Prüfungsteilen sehr uneinheitlich. Die Kandidaten erzielten nicht nur im DSH-Grammatiktest, sondern auch in anderen Prüfungsteilen nur unterdurchschnittliche Ergebnisse. Meistens handelte es sich um Prüfungsteile zum Lesen und/oder zum Schreiben. Bei diesen Kandidaten weisen die Ergebnisse im DSH-Grammatiktest und anderen Prüfungsteilen auf eine Sprachkompetenz, die sich ungleich entwickelte. Die Informationen aus dem DSH-Grammatiktest sind relevant. Sie spiegeln sich allerdings auch in den Ergebnissen in anderen Prüfungsteilen; der DSH-Grammatiktest drückt bestehende Schwierigkeiten besonders deutlich aus.

Bei der Gruppe der "Sprechgewandten 2", welche nur im DSH-Grammatiktest ein unterdurchschnittliches Ergebnis erzielten, stellte sich bei der Betrachtung von Einzelfällen heraus, dass spezifische Schwierigkeiten im Umgang mit den im DSH-Grammatiktest geforderten Fertigkeiten vorlagen. Daraus Rückschlüsse auf die sprachliche Studierfähigkeit zu ziehen, erscheint nach der Betrachtung einzelner Kandidaten als unangemessen.

Bei den "Grammatikfreaks", bei denen der DSH-Grammatiktest eine besonders positive Wirkung auf das Gesamtergebnis hatte, handelt es sich um ausländische Studienbewerber mit ferner Muttersprache. Sieben der elf "Grammatikfreaks" aus der

DSH-TestDaF-Vergleichsstudie stammten aus China bzw. aus Vietnam. Der DSH-Grammatiktest gibt Kandidaten mit fernen Muttersprachen eine Chance zur Ergebnisverbesserung. Mit dem DSH-Grammatiktest können sie im Gesamtergebnis Schwächen im Hörverstehen ausgleichen, für sie ist der Grammatiktest ein "Joker". Dabei muss man beachten, dass die Zahl der Studienbewerber klein sein dürfte, die durch überdurchschnittliche Leistungen im DSH-Grammatiktest Schwächen in anderen Prüfungsteilen ausgleichen können.

Ist es sinnvoll, den "Grammatikfreaks" diese Gelegenheit einzuräumen, oder bietet ein Sprachtest für den Hochschulzugang auch ohne Grammatiktest bereits genügend Informationen zu den Deutschkenntnissen? Ein Grammatiktest ist sinnvoll, wenn die Kandidaten durch die im Grammatiktest gezeigten Fähigkeiten auch dazu befähigt würden, andere Sprachverwendungssituationen im Studium zu meistern. Bei den "Grammatikfreaks" sind die Fertigkeiten Lesen, Sprechen, Schreiben und Hören aber weniger gut ausgebildet. Ich erwarte, dass sie im Fachstudium auf Schwierigkeiten stoßen. Dass sie diese Mängel durch eine überdurchschnittliche Grammatikkompetenz ausgleichen können, halte ich nicht für wahrscheinlich. Insofern ist die Information im Falle der "Grammatikfreaks" zwar interessant, aber nicht zentral für die Zulassungsentscheidung.

Die individuellen Prüfungsergebnisse verdeutlichen, dass der DSH-Grammatiktest nur in Ausnahmefällen die Zulassungsentscheidung beeinflusst. Obwohl der DSH-Grammatiktest einen eigenen Beitrag zum Ergebnis der DSH leistet und die durch den Grammatiktest gewonnenen Informationen nicht durch die übrigen Prüfungsteile abgebildet werden, gab es unter den Kandidaten der DSH-TestDaF-Pilotstudie und der DSH-TestDaF-Vergleichsstudie in Konstanz nur wenige Kandidaten, bei denen der Grammatiktests eine Auswirkung auf die Zulassungsentscheidung hatte. Wenn die Gewichtung weiter reduziert wird, dürfte es kaum Kandidaten geben, bei denen abweichende Leistungen im Grammatiktest das Gesamtergebnis so beeinflusst, dass sich die Ergebnisklasse ändert. Argumente für die Relevanz des DSH-Grammatiktests für die Zulassungsentscheidung konnten aus der Betrachtung individueller Ergebnisse meiner Ansicht nach nicht gewonnen werden.

4.4. Auswirkungen auf Lehr- und Lernprozesse

Überblick: Kapitel 4.4

In diesem Kapitel komme ich zur Frage nach den Auswirkungen des DSH-Grammatiktests auf Lehr- und Lernprozesse. Dazu werden zwei Aspekte erläutert: In einer ersten Teilstudie geht es um die Vorbereitung der Testteilnehmer auf die DSH und die Berücksichtigung der Grammatik dabei. Dazu wurden Umfragen durchgeführt. Als Zweites wird die Behandlung von Grammatik in Lehrbüchern zur Prüfungsvorbereitung analysiert.

4.4.1. Fragestellung und Methode

Sind von einem Grammatiktest als Teil eines Sprachtests für den Hochschulzugang bestimmte Auswirkungen auf die Lehr- und Lernprozesse zu erwarten? Die Begrifflichkeiten und die denkbaren Auswirkungen von Sprachtests auf Lehr- und Lernprozesse wurden in Kapitel 2.2 (Seite 56 f) vorgestellt. Es wurde argumentiert, dass Sprachtests für den Hochschulzugang starke Auswirkungen auf die Prüfungsvorbereitung haben, da sie gewichtige Konsequenzen für die Kandidaten haben. In Kapitel 3.2 (Seite 90 ff) ging ich auf die kontrovers diskutierte Frage ein, welche Rolle die formale Seite der Sprache im Fremdsprachenunterricht einnehmen soll. Ich argumentierte, dass man von negativen Testauswirkungen sprechen sollte, wenn man für einen Sprachtest Fähigkeiten erwerben muss, die man in der realen Sprachverwendungssituation nicht benötigt. Für negativ würde ich es außerdem halten, wenn ein Grammatiktest zu einer isolierten Behandlung von Grammatik in der Testvorbereitung führen oder diese begünstigen würde. Ich befürworte dahingegen einen Unterricht im Sinne von "*focus on form*", bei dem Themen und Aussageabsichten im Mittelpunkt stehen und Grammatik im Dienste der kommunikativen Aussageabsicht vermittelt wird.

Ob Rückwirkungsmechanismen von *deutschen* Sprachtests bislang systematisch untersucht worden sind, ist mir nicht bekannt. Zu anderen Sprachen existiert mittlerweile eine ansehnliche Zahl an Studien (Forschungsübersichten in Alderson/Wall, 1993; Bailey, 1996; Cheng/Watanabe/Curtis, 2004; Wall, 2000). Studien zu den Auswirkungen von Grammatiktests liegen bislang nicht vor. Die Ergebnisse von Untersuchungen zu anderen Tests sind jedoch nicht ohne weiteres zu übertragen. Ist die Studie zu Rückwirkungsmechanismen von japanischen Hochschulzugangsprüfungen, die sich stark von deutschen Prüfungen unterscheiden, für den deutschen Kontext aussagekräftig (Watanabe, 1996)? Kann das Verhalten von zwei Lehrern, die in Hong Kong einen TOEFL-Vorbereitungskurs erteilt haben – und dabei ganz unterschiedlich vorgegangen sind – eine Entscheidungshilfe in Bezug auf die DSH bzw. auf TestDaF sein (Alderson/Hamp-Lyons, 1996)? Relevant dürften allgemeine Erkenntnisse sein: Diese und andere Studien kommen zu dem Schluss, dass sich zwar Rückwirkungsmechanismen beobachten lassen, diese aber bei einzelnen Lehrenden und Lernenden sehr unterschiedlich ausgeprägt sind. Allein die individuelle Vielfalt derjenigen, die am Lehr- und Lernprozess beteiligt sind, verbietet demnach vereinfachende Annahmen über Wirkungen und Ursachen. Schließlich stellt der Rückwirkungseffekt nur einen Faktor im komplexen Bedingungsgefüge von Lehr- Lernprozessen dar. Angesichts der Komplexität wird vor vereinfachenden Ursache-Wirkung-Darstellungen allenthalben gewarnt. Alderson und Wall (1993) ziehen aus ihren umfangreichen Untersuchungen den Schluss, dass Testersteller nicht einfach von bestimmten Auswirkungen ausgehen können, sondern vielmehr in der Pflicht stehen, das Ausmaß und den konkreten Einfluss zu beobachten.

Die empirische Basis zur Beschreibung der konkreten Ausprägung ist dünn, denn beim Nachweis konkreter Auswirkungen einer Prüfung auf die Lehr- und Lernprozesse gibt es methodische Schwierigkeiten: Wie lassen sich Rückwirkungsmechanismen feststellen? Wegen der Komplexität des Phänomens sind aussagekräftige Studien zu Testauswirkungen sehr umfangreich. Studien verfolgen üblicherweise eine allgemeine oder eine spezifische Fragestellung: Wie ändert sich der Unterricht, wenn es am Ende keine Prüfung gibt? Oder: Wie ändert sich der Unterricht/die Prüfungsvorbereitung, wenn die Prüfung geändert wird? Informationen können durch Beobachtungen, Gespräche/ Interviews, mit Fragebögen und Auswertungen von Unterrichtsmaterialien oder Unterrichtsdokumentationen erhoben werden. Diese beziehen üblicherweise Lehrende

und Lernende mit ein (zur Methode siehe Alderson/Wall, 1993; Cheng/Watanabe/ Curtis, 2004).

Obwohl die Auswirkungen eine wichtige Rolle in der Diskussion um den DSH-Grammatiktest darstellen, sind derartige Untersuchungen durch eine Einzelperson nur exemplarisch zu leisten. Die Studien zu den Auswirkungen des DSH-Grammatiktests haben daher einen explorativen Charakter, sie sollen als Ideenbasis für mögliche weitere Untersuchungen zu den Auswirkungen der DSH und des TestDaF auf die Prüfungsvorbereitung dienen.

Fragestellung
Welche Auswirkungen hat der DSH-Grammatiktest auf die Lehr- und Lernprozesse in der Prüfungsvorbereitung?

Ich beschränke die Fragestellung auf die Auswirkungen auf die Lernenden: In welchem Ausmaß wird ihre Prüfungsvorbereitung von dem Grammatiktest beeinflusst? Bei der Leitfrage geht es nicht um die Auswirkungen verschiedener Testmethoden-Merkmale, sondern eher um die Frage, welche Auswirkungen die Existenz bzw. das Fehlen eines Grammatiktests hat.

Informationen zu möglichen Auswirkungen des DSH-Grammatiktests auf die Lehr- und Lernprozesse wurden mit Verfahren erhoben, die sich auf unterschiedliche Aspekte beziehen. Auf diese Weise kann auch mit begrenzten Mitteln ein relativ aussagekräftiger Eindruck entstehen. Zunächst wurden Prüfungskandidaten nach ihrer Prüfungsvorbereitung befragt. Schließlich werden Lehrmaterialien zur Prüfungsvorbereitung auf den TestDaF (ohne Grammatiktest) und auf die DSH (mit Grammatiktest) ausgewertet.

Umfrage zur Studienvorbereitung

Informationen über die Prüfungsvorbereitung wurden mit zwei Umfragen erhoben. Die erste Umfrage bestand aus einem kurzen Fragebogen, den die Kandidaten kurz vor der Durchführung einer DSH an der Fachhochschule Konstanz erhielten. Dieser Fragebogen enthielt Fragen zur Sprachlernbiografie und gezielte Fragen zur Vorbereitung auf die DSH (Abbildung 16).

Haben Sie sich auf einen Prüfungsteil besonders intensiv vorbereitet? Sie können mehrere ankreuzen.

	intensiv vorbereitet	Wenig vorbereitet
Hörverstehen	☐	☐
Textproduktion	☐	☐
Grammatik	☐	☐
Leseverstehen	☐	☐
Mündliche Prüfung	☐	☐

Abbildung 16: Auswirkungen des DSH-Grammatiktests: Umfrage zur Prüfungsvorbereitung

Die zweite Umfrage war Bestandteil der DSH-TestDaF-Vergleichsstudie (siehe Kapitel 4.2.1, Seite 118 ff). Die Kandidaten erhielten einen Fragebogen, der u. a. folgende Frage enthielt:

> "Haben Sie sich auf einen Prüfungsteil besonders gut vorbereitet? (Sie können mehrere Bereiche ankreuzen) Leseverstehen ... "

Mit den beiden Umfragen sollte ermittelt werden, ob die Kandidaten dem Prüfungsteil "Wissenschaftssprachliche Strukturen" im Rahmen ihrer Prüfungsvorbereitung eine besondere Bedeutung zumaßen. Anhand des Ergebnisses aus dem Grammatiktest kann auch ermittelt werden, wie erfolgreich die Vorbereitung war.

Die Ergebnisse beider Umfragen werden in Kapitel 4.4.2 (Seite 163 ff) dargestellt und diskutiert.

Auswertung von Lehrmaterialien

Materialien für die Prüfungsvorbereitung stehen in einem engen Zusammenhang mit Prüfungsmerkmalen. Interessant ist der Vergleich von Lehrbüchern zur Vorbereitung auf die DSH mit Lehrbüchern zur Vorbereitung auf den TestDaF. Da der TestDaF im Gegensatz zur DSH keinen expliziten Grammatiktest enthält, lässt sich der Einfluss eines solchen Prüfungsteils auf die Rolle der Grammatik in Lehrbüchern zur gezielten Prüfungsvorbereitung ermitteln. Es werden alle aktuellen Lehrbücher zur Vorbereitung auf die DSH sowie auf den TestDaF berücksichtigt.

In der Regel geht man von einem großen Zusammenhang zwischen Lehrwerk und Unterricht aus (Hüllen/Löscher, 1979; Kast/Neuner, 1994). In der "*Sri Lankan impact study*", welche die Einführung von neuen Prüfungen und neuen Lehrmaterialien in Sri Lanka begleitete, wurde diese Annahme genauer untersucht (Wall/Alderson, 1995). In der umfangreichen Studie wurden die Auswirkungen der neuen Prüfungen (und der neuen Lehrbücher) anhand einer Vielzahl von Faktoren erhoben. Es wurden beispielsweise nicht nur die Lehrbücher zur Vorbereitung auf die neuen Prüfungen untersucht, sondern es wurde auch der Unterricht mit den Lehrbüchern beobachtet. Man verglich dabei den Anteil, den Übungen für eine bestimmte Fertigkeit im Buch einnahmen, mit dem Anteil, der im Unterricht zur Übung der Fertigkeit eingeräumt wurde. Dabei traf man auf große Unterschiede. Es wurde außerdem beobachtet, dass viele Lehrkräfte lieber auf eigene Materialien zurückgriffen. Für die Studien zum DSH-Grammatiktest sind die Ergebnisse aus der "*Sri Lankan impact study*" sicherlich nicht ohne weiteres übertragbar. Die Studie wurde von mir angeführt, weil sie exemplarisch einen methodischen Weg zur Erhebung von Testauswirkungen beschreitet.

Auf die Schwierigkeiten, allgemein gültige Hinweise zu den Auswirkungen von Prüfungen zu gewinnen, wurde bereits hingewiesen. Die Auswertung von einigen Lehrbüchern kann lediglich als ein Indiz gewertet werden. Lehrbücher sind nur einer von mehreren Faktoren, welche bei einer umfassenderen Betrachtung der Prüfungsauswirkungen zu berücksichtigen sind.

Die Auswertung von Lehrmaterialien erfolgt in Kapitel 4.4.3 (Seite 167 ff).

4.4.2. Auswirkungen auf Lehr- und Lernprozesse: Ergebnisse und Diskussion

Ergebnisse aus einer Umfrage zur Prüfungsvorbereitung

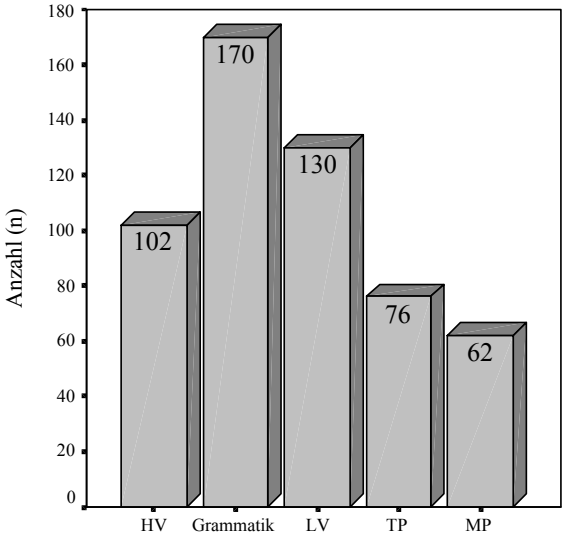
Insgesamt 305 Kandidaten beantworteten die Fragen zu ihrer Vorbereitung auf die DSH. In der Vorbereitung wird dem DSH-Grammatiktest demnach eine große Aufmerksamkeit geschenkt. 170 Kandidaten (56 Prozent) gaben an, sie hätten sich auf den Grammatiktest besonders intensiv vorbereitet. Mit einigem Abstand folgen die Prüfungsteile Leseverstehen, Hörverstehen, Textproduktion und schließlich die Mündliche Prüfung. Die Verteilung der Häufigkeiten geht aus der Tabelle 25 und der Abbildung 17a (Seite 165) hervor.

Die häufige Konzentration der Kandidaten auf den DSH-Grammatiktest wird besonders deutlich, wenn nur die Kandidaten in die Analyse einbezogen werden, die sich auf einen oder mehrere Prüfungsteile besonders intensiv vorbereitet haben, nicht aber gleichmäßig auf alle Prüfungsteile oder auf keinen Prüfungsteil (siehe Zeile "auf einen bis vier Prüfungsteile vorbereitet", Tabelle 25; Abbildung 17b, Seite 165). Nicht berücksichtigt wurden also weder die 73 Kandidaten (24 Prozent), die angaben, sich nicht gezielt auf einen oder mehrere Prüfungsteile vorbereitet zu haben, noch die 41 Kandidaten (13 Prozent), welche der Meinung waren, sich auf alle Prüfungsteile gleichmäßig vorbereitet zu haben. Von den übrigen 191 Kandidaten, die sich gezielt auf einen oder mehrere Prüfungsteile vorbereiteten, zählten 129 (68 Prozent) den DSH-Grammatiktest dazu. Mit deutlichem Abstand folgen die Prüfungsteile Leseverstehen (47 Prozent), Hörverstehen (32 Prozent), Textproduktion (18 Prozent) und Mündliche Prüfung (11 Prozent). Das heißt, wenn sich die Kandidaten in ihrer Prüfungsvorbereitung gezielt auf einen oder mehrere Prüfungsteile konzentrierten, gehörte der DSH-Grammatiktest in zwei Dritteln der Fälle dazu. Die rezeptiven Prüfungsteile Leseverstehen und Hörverstehen finden eine mittlere Beachtung in der Prüfungsvorbereitung. Eher vernachlässigt werden die produktiven Prüfungsteile Textproduktion und Mündliche Prüfung in der Vorbereitung der Kandidaten.

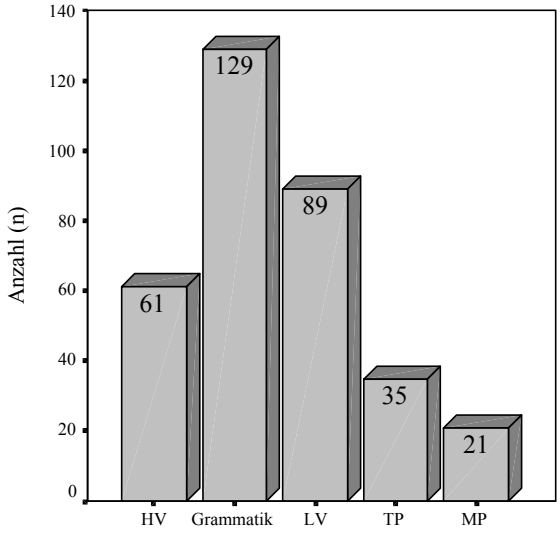
99 Kandidaten gaben an, sie hätten sich besonders auf einen einzigen Prüfungsteil konzentriert (siehe Zeile "nur auf einen Prüfungsteil vorbereitet", Tabelle 25; Abbildung 17c). Betrachtet man die Rückmeldung dieser Prüflinge, so wird die Bedeutung des DSH-Grammatiktests noch einmal besonders greifbar. 59 Prozent konzentrierten sich in der Prüfungsvorbereitung auf den DSH-Grammatiktest. Mit großem Abstand folgen die rezeptiven Prüfungsteile Leseverstehen und Hörverstehen (je 15 Prozent). Wenn sich die Kandidaten in der Vorbereitung auf einen einzigen Prüfungsteil konzentrierten, gehörten die produktiven Prüfungsteile in der Regel nicht dazu.

Tabelle 25: DSH – Besondere Vorbereitung auf Prüfungsteile

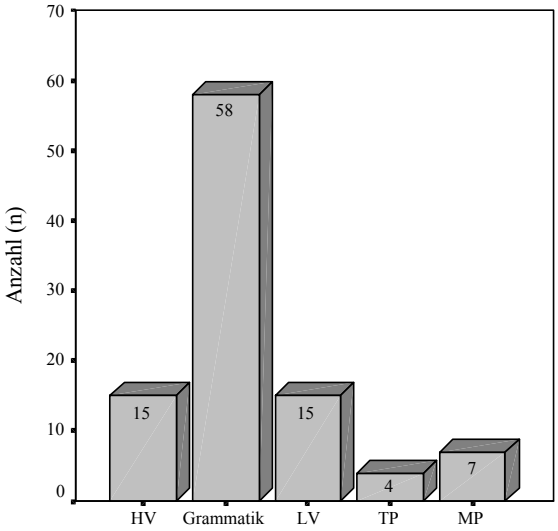
	Hör-verstehen	**Gram-matik**	**Lese-verstehen**	**Text-produktion**	**Mündliche Prüfung**
alle Kandidaten (n = 305)	102 33 %	170 56 %	130 43 %	76 25 %	62 20 %
auf einen bis vier Prüfungs-teile vorbereitet (n = 191)	61 32 %	129 68 %	89 47 %	35 18 %	21 11 %
nur auf einen Prüfungsteil vorbereitet (n = 99)	15 15 %	58 59 %	15 15 %	4 4 %	7 7 %



a) Vorbereitung auf ...



b) Vorbereitung auf einen bis vier Prüfungsteile



c) Vorbereitung auf einen Prüfungsteil, nämlich ...

Abbildung 17: DSH – Besondere Vorbereitung auf Prüfungsteile (Säulendiagramme)

Ergebnisse aus der DSH-TestDaF-Vergleichsstudie

Die Ergebnisse aus der DSH-TestDaF-Vergleichsstudie bestätigen die Beobachtungen. Der Grammatiktest ist wiederum der Prüfungsteil, der in der Vorbereitung der Kandidaten die größte Aufmerksamkeit auf sich zieht (siehe Tabelle 26). Wegen der kleineren Anzahl der Kandidaten sind die Ergebnisse allerdings weniger aussagekräftig.

Viele Prüfungskandidaten sind offensichtlich der Meinung, dass der Prüfungsteil Grammatik einer bestimmten Vorbereitung bedarf, während sie hoffen, die produktiven Prüfungsteile irgendwie schon zu bestehen. Der Grammatiktest scheint die Prüfungsvorbereitung der Kandidaten zu dominieren. Viele Kandidaten hatten zur Vorbereitung auf die DSH einen Sprachkurs besucht. Eine ebenfalls große Gruppe bereitete sich alleine auf die DSH vor. Diese Kandidaten konzentrierten sich auf den Grammatiktest. Wenn die Kandidaten sich in einem Sprachkurs vorbereiteten, war die Konzentration auf nur einen Prüfungsteil seltener. Offensichtlich wird der Prüfungsteil Grammatik von den Kandidaten als besonders wichtig angesehen. In Sprachkursen scheint sich die Vorbereitung nicht allein auf den Grammatiktest zu konzentrieren.

Tabelle 26: DSH-TestDaF-Vergleichsstudie – Besondere Vorbereitung auf Prüfungsteile

	Hör-verstehen	**Gram-matik**	**Lese-verstehen**	**Text-produktion**	**Mündliche Prüfung**
alle Kandidaten (n = 56)	26 46 %	38 68 %	29 52 %	21 38 %	15 27 %
auf einen bis vier Prüfungsteile vorbereitet (n = 43)	19 44 %	31 72 %	22 51 %	14 33 %	8 19 %
nur auf einen Prüfungsteil vorbereitet (n = 13)	3 23 %	8 62 %	1 8 %	1 8 %	0 0 %

4.4.3. Auswertung von Lehrmaterialien

Führt der DSH-Grammatiktest dazu, dass Grammatik in Lehrmaterialien zur Vorbereitung auf die DSH stärker berücksichtigt wird als in Lehrmaterialien zur Vorbereitung auf den TestDaF? Hinweise sollen aus sechs Lehrbüchern zur unmittelbaren Prüfungsvorbereitung gewonnen werden (siehe Tabelle 27). In diesen Lehrbüchern werden die Kandidaten mit dem Format der jeweiligen Prüfung vertraut gemacht. Sie enthalten auch Musterprüfungen. Alle Materialien eignen sich laut Klappentext sowohl für den Selbstlerneinsatz als auch für den Unterrichtseinsatz. Die Behandlung der Grammatik unterscheidet sich jedoch.

Tabelle 27: Grammatik in Lehrbüchern zur Vorbereitung auf die DSH bzw. auf den TestDaF

Titel	Test	Vermittlung von Grammatik
Lodewick, K. 2001. DSH & Studienvorbereitung. Vorbereitung auf ein Studium an einer deutschsprachigen Universität. Göttingen: Fabouda.	DSH	ja
Eggers, D. Müller-Küppers, E.; Wiemer, C.; Zöllner, I. 1999. Prüfungskurs DSH. Vorbereitung auf die Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber. Ismaning: Hueber.	DSH	ja
Jung, L. 1995. Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber (DSH). Ismaning: Hueber.	DSH	ja
Gutzat, B.; Kniffka, G.. 2003. Training TestDaF – Material zur Prüfungsvorbereitung. Berlin und München: Langenscheidt.	TestDaF	nein
Lodewick, K. 2002. TestDaF-Training. Text- und Übungsbuch zur Vorbereitung auf den TestDaF. Göttingen: Fabouda.	TestDaF	nein
Glienicke, S.; Katthagen, K.-M. 2003. TestDaF – Kurs zur Prüfungsvorbereitung. Ismaning: Hueber.	TestDaF	ja

Lodewick, K. (2001).* DSH & Studienvorbereitung. Vorbereitung auf ein Studium an einer deutschsprachigen Universität. *Göttingen: Fabouda Verlag.

In diesem Buch sind die Aufgaben jeweils einem DSH-Prüfungsteil zugeordnet; es enthält also auch Aufgaben zum Grammatiktest. Dennoch handelt es sich nicht um eine

Sammlung von Musterprüfungen. Es werden Aufgaben und Übungen angeregt, die über die Prüfungsaktivitäten hinausgehen. Zu den Lesetexten werden beispielsweise Aufgaben zur aktiven Vorentlastung der Texte gestellt. Die prüfungsrelevanten Aufgabentypen werden analysiert und schrittweise bearbeitet. So findet auch eine Erweiterung der methodischen Kompetenzen statt. Es werden unterschiedliche Grammatiktests vorgestellt: Es finden sich Grammatiktests, die ähnlich wie der Grammatiktest "Teilzeitarbeit" (siehe Abbildung 10, Seite 103) Metasprache verwenden, wie auch Grammatiktests, die auf metasprachliche Anweisungen verzichten, wie im Grammatiktest "Flurbereinigung" (siehe Abbildung 7, Seite 74). Zum Grammatiktest werden hier auch textuelle Referenzen und Worterklärungen gezählt. Eine dekontextualisierte Konzentration auf Grammatikphänomene findet nicht statt.

Eggers, D.; Müller-Küppers, E.; Wiemer, C.; Zöllner, I. (1999). **Prüfungskurs DSH. Vorbereitung auf die Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber.** ***Ismaning: Hueber.***

Das Buch enthält zwei Hauptteile. Im ersten wird jeder Prüfungsteil mit einem ausführlichen Beispiel erläutert und es werden Hinweise zur Beantwortung gegeben. Es gibt also ein eigenes Kapitel für den DSH-Grammatiktest, in dem die Kandidaten mit möglichen Aufgabentypen vertraut gemacht werden. Im zweiten Hauptteil werden fünf Grammatiktests zur Übung angeboten, die sich jeweils auf Lesetexte beziehen.
Beispiele für Aufgabenstellungen sind:

- "Worauf bezieht sich 'alle' (Zeile ...)?"
- "Warum wird in Zeile ... der Konjunktiv II benutzt?"
- "Schreiben Sie die folgenden Sätze um, indem Sie die in Klammern angegebenen Strukturen verwenden" (z. B. Partizipialattribut).

Auch hier werden die Kandidaten auf die Verwendung von Metasprache vorbereitet. Ein inhaltsleeres Abarbeiten der Grammatikaufgaben ist nicht möglich. Das liegt nicht zuletzt daran, dass etwa die Hälfte der Aufgaben eher dem Prüfungsteil Textverstehen zuzurechnen ist.

Jung, L. (1995). **Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber (DSH).** ***Ismaning: Hueber.***

Dieses Lehrbuch wurde vor der Erstellung des DSH-Handbuchs veröffentlicht. Jedem Prüfungsteil der Schriftlichen Prüfung ist ein Kapitel gewidmet. Allerdings finden sich keine Aufgaben zum Prüfungsteil Textproduktion. Das Kapitel "Wissenschaftssprachliche Strukturen" enthält 15 typische DSH-Grammatiktests, die sich inhaltlich an Lesetexten des folgenden Kapitels ("Bearbeitung von Lesetexten") orientieren. Die Lerner erhalten einen Schlüssel zu den Übungen, aber keine Hinweise zur Lösung. Typische Aufgabenstellungen sind:

- "Formen Sie die Gliedsätze in Satzglieder um."
- "Ersetzen Sie den jeweils kursiv gedruckten Ausdruck durch ein passendes Modalverb (können, mögen, müssen, wollen)."
- "Formen Sie die Attributsätze (Relativsätze) in Partizipialattribute um."
- "Ergänzen Sie die Präpositionen."

Es werden ausschließlich Aufgaben mit Metasprache verwendet, die häufig auch geordnet dargeboten werden wie im Grammatiktest Ordnung "Neue Medien" (siehe Abbildung 11, Seite 103).

Im Informationsteil wird die Arbeit in Gruppen und eine Bearbeitung mit einem verkürzten Zeitlimit angeregt. Außerdem wird die Arbeit mit einer Grammatik sowie die "Aufarbeitung eines Phänomens" im Unterricht empfohlen (Jung, 1995: 8). Eine Anleitung dazu ist nicht Teil des Lehrbuchs. Das Lehrbuch von Jung dient im Sinne eines Übungsrepititoriums ausschließlich der Vorbereitung auf das Format der DSH.

Gutzat, B.; Kniffka, G. (2003). **Training TestDaF – Material zur Prüfungsvorbereitung: Trainingsbuch.** ***Berlin und München: Langenscheidt.***

Der Aufbau des Lehrbuchs "Training TestDaF" orientiert sich an den Prüfungsteilen des TestDaF. Es besteht folgerichtig aus vier Kapiteln: "Training Leseverstehen", "Training Hörverstehen" usw. Es enthält außerdem einen Modelltest und die Lösungen zu den Aufgaben. Zwischen den Kapiteln wird keine Progression verfolgt, so dass die Reihenfolge, in der die Kapitel bearbeitet werden, unerheblich ist. Innerhalb der einzelnen Kapitel ist eine sprachliche Progression zu beobachten, was auch an dem unterschiedlichen Schwierigkeitsgrad der Lese- bzw. Hörverstehensaufgaben des TestDaF liegt, die

nachvollzogen wird. Das Lehrbuch besteht jedoch nicht nur aus Musterprüfungen, jeder Prüfungsteil wird vielmehr sehr detailliert eingeführt. Die Anweisungen zum Prüfungsteil "Schriftlicher Ausdruck" enthalten beispielsweise Redemittel ("Verben, die einen Rückgang bezeichnen: abnehmen/fallen/sinken/...", Gutzat/Kniffka, 2003: 60), Tipps für die Gliederung von Texten ("Ihr Prüfungstext sollte die folgenden Gliederungspunkte enthalten", 54) sowie Wortschatzübungen ("Ergänzen Sie im folgenden Text die unterstrichene Wörter durch sinnverwandte", 62). Grammatische Strukturen werden nicht explizit vermittelt oder neu eingeführt. Gleichwohl werden Fertigkeiten geübt, die auch im Grammatiktest geprüft werden: "Verknüpfen sie die folgenden Feststellungen/ Forderungen und Begründungen. Überlegen Sie sich jeweils möglichst mehrere Varianten" (Gutzat/Kniffka, 2003: 63). Sie stehen aber stets im Dienst der im jeweiligen TestDaF-Prüfungsteil geprüften Fertigkeit. Es wird vorausgesetzt, dass die sprachlichen Strukturen, die für die Erfüllung der Aufgabe erforderlich sind, bereits bekannt sind. Geübt werden die Anwendung und die inhaltliche Auseinandersetzung mit dem Thema.

Lodewick, K. (2002). **TestDaF-Training. Text- und Übungsbuch zur Vorbereitung auf den TestDaF.** ***Göttingen: Fabouda Verlag.***

Es überrascht nicht, dass ein Lehrbuch mit Titel "TestDaF-Training" ähnlich vorgeht wie "Training TestDaF". Das Text- und Übungsbuch besteht aus drei Teilen: Einer Anleitung zum Umgang mit den vier Prüfungsteilen des TestDaF, einem Teil "Üben und Trainieren" und einem Modelltest. Auch im Hauptteil "Üben und Trainieren", der den größten Raum einnimmt, ist jeder Abschnitt einem TestDaF-Prüfungsteil zugeordnet. Eine explizite Vermittlung von Grammatik findet nicht statt. Das Buch enthält einige Aufgaben zu sprachlichen Strukturen, die jeweils im Dienst eines der vier TestDaF-Prüfungsteile stehen: Die Auflösung von Komposita ("Bedienungsanleitung = eine ... für die ... eines Gerätes", Lodewick, 2002: 59) steht beispielsweise im Zusammenhang mit dem Leseverstehen, Hinweise zu Nebensätzen ("während + Nebensatz", 37) stehen im Zusammenhang mit der Textproduktion. Das Buch konzentriert sich auf die unmittelbaren Anforderungen des TestDaF. Im Gegensatz zu "DSH & Studienvorbereitung" vom selben Autor wird die Grammatik konsequenterweise nicht thematisiert.

"TestDaF-Training" und "Training TestDaF" sind Beispiele für prüfungsvorbereitende Lehrbücher, die sich ausschließlich auf die Prüfung konzentrieren. Das übergeordnete

Ziel, die sprachliche Vorbereitung auf ein Studium wird nur insofern gefördert, als dass die im TestDaF geforderten Fertigkeiten eine Nähe zu den sprachlichen Fertigkeiten haben, die im Studium benötigt werden. Folgerichtig wird Grammatik nicht explizit vermittelt und geübt.

Glienicke, S.; Katthagen, K.-M. (2003).* TestDaF – Kurs zur Prüfungsvorbereitung. *Ismaning: Hueber.

Das Lehrbuch enthält zwar auch eine TestDaF-Modellprüfung und viele Hinweise zum TestDaF; im Gegensatz zu den anderen beiden Lehrbüchern zum TestDaF wurde es jedoch tatsächlich als kurstragendes Unterrichtsbuch mit einer sprachlichen Progression konzipiert. Strukturelle Phänomene werden explizit thematisiert, obwohl sie im TestDaF nicht explizit abgefragt werden. In den einführenden Hinweisen zum Lehrbuch nehmen die Autoren Stellung zur Behandlung von Grammatik:

> Grammatik und Strukturübungen haben im Prüfungscurriculum des TestDaF gegenüber der Textarbeit keine unmittelbare Bedeutung, sie werden jedoch erfahrungsgemäß von den Studierenden stark nachgefragt und dienen der mittelbaren Vorbereitung des Prüfungskönnens. Das Buch versucht diesen Anforderungen insofern gerecht zu werden, als es systematisch und progressiv aufgebaute Übungen zum Grammatikgerüst anbietet. Eine weiter gehende, vertiefende Einübung des jeweiligen Grammatik-Schwerpunkts bleibt jedoch dem empfohlenen Begleitmaterial vorbehalten (Glienicke/Katthagen, 2003: 7).

Bei dem Begleitmaterial handelt es sich um drei Grammatikbücher, auf die bei mehreren Übungen verwiesen wird.

Die Grammatikübungen folgen einer Progression, wobei einmal thematisierte Phänomene in späteren Kapiteln wieder aufgegriffen werden. Zu den behandelten Themen zählen: Indirekte Rede, Konjunktiv I und II, zusammengesetzte Substantive, Funktionsverbgefüge, Partizipien als Attribute, Präpositionen usw. Auffällig ist, dass die Vermittlung von Metasprache als Kursziel angesehen wird. Daher sind Aufgaben enthalten wie: "Finden Sie im Text je drei Beispiele für: Zusammengesetzte Substantive, Nominalisierungen, Nomen-Verb-Verbindungen" (Glienicke/Katthagen, 2003: 104). Da keine Beispiele gegeben werden, kann diese Aufgabe nur lösen, wer mit den Fachbegriffen vertraut ist. Im Inhaltsverzeichnis findet sich dazu die Überschrift: "Grammatik metasprachlich entwickeln".

Das Lehrbuch von Glienicke und Katthagen ist ein deutlicher Hinweis darauf, dass der Verzicht auf einen Grammatiktest in einem Sprachtest für den Hochschulzugang nicht zwangsläufig zu einem Verzicht auf Grammatikvermittlung führt. Interessant ist der

Hinweis der Autoren auf das Interesse der Kursteilnehmer an der Grammatikvermittlung. In der Interessenlage der Kursteilnehmer liegt ihrer Ansicht nach ein Hauptargument für den Fortbestand der Grammatikvermittlung. Begründet werden die Grammatikübungen nicht mit dem Wert für den TestDaF oder für die Förderung der studienrelevanten Deutschkompetenz.

Die Existenz unterschiedlicher Lehrbücher zur Prüfungsvorbereitung auf die DSH und den TestDaF deutet darauf hin, dass die Unterschiede zwischen den Prüfungen sowohl von den Autoren als auch von den ausländischen Studienbewerbern wahrgenommen werden. Offensichtlich gehen Autoren und Verlage davon aus, dass es sich lohnt, unterschiedliche Titel zu verlegen, da sich diese besser verkaufen lassen als ein Titel, der auf beide Prüfungen vorbereitet. Bemerkenswert ist, dass Grammatik in den Lehrbüchern zur Vorbereitung auf den TestDaF unterschiedlich behandelt wird. In einem Buch wird Grammatik vermieden, in einem anderen bewusst vermittelt (siehe Tabelle 27, Seite 167).

Die Lehrbücher zum TestDaF können sich wegen der Standardisierung der Prüfung wesentlich genauer auf das Format beziehen als die Lehrbücher zur DSH. Alle Lehrbücher zum TestDaF enthalten daher auch Musterprüfungen. Die Lehrbücher zur DSH müssen auf eine Prüfungsvielfalt vorbereiten. Dies wird bisweilen durch die Verwendung verschiedener Formate anerkannt (z. B. Lodewick, 2001), bisweilen aber auch ignoriert (z. B. Jung, 1995).

Die Botschaft der Kontextualisierung der DSH-Grammatikaufgaben scheint bei den Autorinnen und Autoren von Lehrmaterialien zur Prüfungsvorbereitung angekommen zu sein. Sie beziehen den Grammatiktest ein, ohne jedoch die Strukturen dekontextualisiert zu vermitteln. Ganz ohne eine besondere Vorbereitung auf den DSH-Grammatiktest – so die Wahrnehmung der Lehrbuchautoren – geht es wohl nicht.

4.4.4. Zusammenfassung und Diskussion

Fragestellung
Welche Auswirkungen hat der DSH-Grammatiktest auf die Lehr- und Lernprozesse in der Prüfungsvorbereitung?

DSH-Prüfungskandidaten bereiteten sich häufig gezielt auf den DSH-Grammatiktest vor. Dies ergab eine Befragung von 305 Teilnehmern an einer DSH. Wie kommt es zu der Konzentration auf den DSH-Grammatiktest? Mehrere Ursachen sind denkbar: Möglicherweise versprechen sich die Kandidaten von einem Lerneinsatz für den Grammatiktest eine große Wirkung. Vielleicht eignet sich der Grammatiktest aber auch zur Vorbereitung, weil er als einziger DSH-Prüfungsteil "lernbares" Wissen abfragt. Möglicherweise ist die Vorbereitung auf den Grammatiktest offensichtlicher und besser allein zu leisten. Für diese Argumentation spricht, dass sich vornehmlich Kandidaten auf den Grammatiktest vorbereiteten, welche nicht an Sprachkursen teilnahmen.

Die Berücksichtigung von Grammatik in Materialien zur Vorbereitung auf die DSH bzw. den TestDaF ist uneinheitlich. Es ist typisch für Untersuchungen der Rückwirkungsmechanismen von Sprachtests, dass man auf uneinheitliche Auswirkungen trifft: Während ein Lehrbuch zur Prüfungsvorbereitung auf den TestDaF auf Grammatik verzichtet, wird sie in einem anderen bewusst vermittelt. So dürfte es auch im Unterricht sein: Während für einige Lehrkräfte der Verzicht auf einen Grammatiktest im TestDaF ein Anlass ist, den Anteil der Grammatikvermittlung zu reduzieren oder gar zu streichen, behalten andere eine explizite Grammatikvermittlung bei. Lehrbücher zur Vorbereitung auf die DSH enthalten Aufgaben zum Grammatiktest, in einigen wird aber auf Erklärungen verzichtet.

Die eingesetzten Methoden waren unterschiedlich produktiv. Die Studien zu den Auswirkungen des DSH-Grammatiktests wurden auch mit der Zielsetzung durchgeführt, Methoden zur Erhebung von Testauswirkungen zu evaluieren. Als produktiv erwies sich meiner Ansicht nach die Befragung der Testteilnehmer nach ihrer Vor-

bereitung. An Lehrwerken lassen sich konkrete Auswirkungen aufzeigen. Ob die Lehrwerke auch die Unterrichtswirklichkeit spiegeln, ist jedoch eine offene Frage.

Lassen sich aus den Studien zu den Auswirkungen des DSH-Grammatiktests Argumente für oder gegen die Legitimität dieses Prüfungsteils ableiten? Die Beantwortung der Frage hängt letztendlich von dem Bewertungsrahmen ab: ob man nämlich eine Konzentration auf die formale Seite der Sprache im studienvorbereitenden Sprachenunterricht für sinnvoll hält oder nicht. In Kapitel 3.2 (Seite 90 ff) sprach ich mich für einen Unterricht aus, der im Sinne von *focus on form* Inhalte in den Mittelpunkt und die Vermittlung von Strukturen in den Dienst der Mitteilungsabsichten stellt. Die starke Konzentration auf den DSH-Grammatiktest in der Studienvorbereitung deutet meiner Ansicht nach jedoch darauf hin, dass es auch um *focus on formS* geht, dass möglicherweise also auf eine inhaltliche Anbindung verzichtet wird. Dies halte ich nicht für sinnvoll.

4.5. Grammatik in Sprachtests für den Hochschulzugang: Ausblick

Übersicht: Kapitel 4.5
Im letzten Kapitel zu Grammatik in Sprachtests für den Hochschulzugang fasse ich die Ergebnisse der beschriebenen Studien zur Reliabilität des DSH-Grammatiktests, zur Konstruktvalidität und zu den Auswirkungen zusammen und ziehe Schlussfolgerungen daraus.

Am DSH-Grammatiktest, der häufig wie eine Kursabschlussprüfung eingesetzt wird, manifestiert sich eine dezentrale, unterrichtsorientierte Prüfungstradition. Kandidaten, die nicht an speziellen Prüfungsvorbereitungskursen teilgenommen haben, dürften benachteiligt sein. Vorgesehen ist laut DSH-Handbuch ein indirekter Kompetenztest, der mittels kontextualisierter Transfomationsaufgaben produktive Grammatikkompetenz ohne den Einsatz von Metasprache und ohne Ordnung der Phänomene prüft.

Die DSH leidet an der fehlenden Standardisierung: Wie eng sich einzelne Ausrichter an den Vorgaben der Rahmenordnung oder des DSH-Handbuchs orientieren, ist nicht bekannt. Daher wurden auch Grammatiktests in die Studien einbezogen, welche vom Format der "DSH-Prototyp-Grammatik" abweichen (Kapitel 4.1). Es wurde festgestellt, dass unterschiedliche Formate durchaus zu unterschiedlichen Ergebnissen führen. Die Paralleltestreliabilität von Grammatiktests mit unterschiedlichen Testmethoden-Merkmalen ist nicht hoch genug; von einer Äquivalenz der Tests konnte keine Rede sein. Die Reliabilität könnte etwas erhöht werden, wenn alle Anbieter den DSH-Grammatiktest nach den gleichen Testmethoden-Merkmalen erstellen. Dies ist jedoch derzeit nicht gesichert.

Eine Verwendung von Metasprache oder die Ordnung der sprachlichen Phänomene erwies sich als ungünstig: Bei fehlerhaften Antworten bleibt unklar, ob die Metasprache

oder das sprachliche Phänomen unbekannt ist. Derartige Testmethoden-Merkmale führen zu einer Benachteiligung von Kandidaten, die nicht am Vorbereitungskurs teilnahmen.

In der Studie differenzierten Grammatiktests, die nach Vorgabe des DSH-Handbuchs konzipiert wurden, stärker als andere DSH-Prüfungsteile zwischen den Leistungen der Kandidaten. Die Fähigkeit, mit Schriftsprache umzugehen, korrelierte stark mit den Fähigkeiten, die im DSH-Grammatiktest benötigt werden; der Umgang mit gesprochener Sprache beruht jedoch auf davon zu unterscheidenden Fertigkeiten, er wird durch den DSH-Grammatiktest nicht oder nur am Rande erfasst (Kapitel 4.2). Fest steht, dass der DSH-Grammatiktest einen Einfluss auf das Konstrukt der Schriftlichen Prüfung hat. Ein Verzicht auf den Prüfungsteil zur Grammatik würde zu einem Verlust an Informationen führen: Prüfungsteile zum Schreiben und z. T. auch zum Lesen erwiesen sich zwar als signifikante Prädiktorvariablen für die Ergebnisse des DSH-Grammatiktests, sie können jedoch einen großen Anteil der Variation statistisch nicht erklären. Im Sinne von Purpuras (2004) Definition von Grammatikkompetenz gehe ich davon aus, dass es sich dabei um produktive Grammatikkompetenz handelt.

Ob die zusätzlichen Informationen zur produktiven Grammatikkompetenz, welche der DSH-Grammatiktest ermittelt, für die Entscheidung über die Zulassung zum Studium von Belang sind, ist fraglich. Die Informationen, die ohne einen Grammatiktest gewonnen werden können, dürften für die Entscheidung ausreichen (Kapitel 4.3). Nur eine kleine Gruppe ausländischer Studienbewerber profitiert von der Existenz des DSH-Grammatiktests im Sinne eines "Prüfungsjokers". Dabei scheint es sich häufig um Kandidaten mit einer fernen Muttersprache zu handeln. Die sprachliche Richtigkeit im Umgang mit Schriftsprache ist für sie eher zu bewältigen als der Umgang mit gesprochener Sprache. Wenn allein das Ergebnis im DSH-Grammatiktest auf sprachliche Defizite hindeutet, ist dies für die Zulassungsentscheidung nicht relevant: Es ist unerheblich, ob Kandidaten, die im Schreiben, Sprechen, Lesen und Hören Sprachkenntnisse auf Oberstufenniveau zeigen, Schwierigkeiten mit den Transformationsaufgaben des DSH-Grammatiktests haben.

Welche Auswirkungen hat der DSH-Grammatiktest auf Lehr- und Lernprozesse (Kapitel 4.4)? Die Existenz des Grammatiktests führt bei einer großen Zahl von Kandidaten zu einer intensiven Auseinandersetzung mit der strukturellen Seite der

Sprache. Die Umfrage zur Prüfungsvorbereitung legt außerdem nahe, dass dies vor allem auf Kandidaten zutrifft, welche sich nicht in Sprachkursen, sondern allein auf die DSH vorbereiten. In Lehrbüchern zur Vorbereitung auf die DSH, welche als Indikator für das Vorgehen im Unterricht interpretiert wurden, spielt die Grammatik selbstredend eine wichtige Rolle. In Lehrbüchern zur Vorbereitung auf den TestDaF wird die strukturelle Seite der Sprache nur in Ausnahmefällen explizit vermittelt.

Die Legitimität des DSH-Grammatiktests hängt damit zusammen, wie wichtig sprachliche Richtigkeit für ausländische Studierende ist und wie viel Wert daher auf das Training der sprachlichen Richtigkeit gelegt werden soll. Zu dieser Frage gibt es verschiedene Ansichten, und es liegt in der Natur der Sache, dass sie weniger empirisch abgesichert sind, als vielmehr ein Ergebnis subjektiver Theorien und Erfahrungen darstellen. In Kapitel 2.2 (Seite 56) wurde argumentiert, dass Auswirkungen als positiv zu betrachten sind, wenn sich die Lerner Fähigkeiten aneignen, die sie nicht nur im bevorstehenden Sprachtest, sondern auch in der angestrebten Sprachverwendungssituation einsetzen können. Erwähnt wurde auch Messicks Anregung, sich bei der Testkonstruktion um die Abbildung des sprachlichen Konstrukts zu bemühen und auf diese Weise eine hohe Validität zu gewährleisten. Die im DSH-Grammatiktest geforderten Transformationsaufgaben werden von ausländischen Studierenden im Studium in dieser Form nicht verlangt. Sinnvoll ist ein Training derartiger Umformungen nur, wenn die Beschäftigung mit den geforderten Strukturen den Kandidaten in Sprachverwendungssituationen im Studium weiterhilft, wenn sie damit etwa Texte besser verstehen, in Klausuren besser formulieren, Vorlesungen besser verstehen oder besser an Diskussionen teilnehmen können. Nur wenn die im DSH-Grammatiktest geforderten Fähigkeiten also ein Bestandteil der "ausreichenden Deutschkenntnisse für das Studium" sind und mithin ein Indikator dafür, ist es legitim, von Kandidaten, die Maschinenbau, Jura oder Medizin studieren möchten, sprachliche Aktivitäten zu verlangen, auf die sie im Studium nicht treffen.

Vor dem Hintergrund der in Kapitel 4.4 gezeigten starken Konzentration der Prüfungskandidaten auf den Grammatiktest halte ich es für angemessen, dem Grammatiktest innerhalb der DSH ein geringeres Gewicht beizumessen oder ihn ganz zu streichen. Möglicherweise führen ein reduzierter Umfang und eine geringere Gewichtung zu einer

ausgeglicheneren Vorbereitung. Auch eine Angliederung an das Leseverstehen, wie in der aktuellen Rahmenordnung vorgesehen, ist möglich.

Diese Empfehlungen wurden im Wesentlichen in der DSH-Rahmenordnung von 2004 umgesetzt. Allerdings ist in der Rahmenordnung nicht gelungen, eine stimmige Konzeption zu erstellen: In Kapitel 2.2 (Seite 59 ff) wurde bereits darauf hingewiesen, dass es kaum zu vermitteln ist, wenn eine Prüfung aus drei Prüfungsteilen mit vier Ergebnissen besteht. "Leseverstehen und wissenschaftssprachliche Strukturen" ist *ein* Prüfungsteil, bei dem beide Teile allerdings *einzeln* bewertet und einzeln im Zeugnis aufgeführt werden. Beide Teile haben außerdem unterschiedliches Gewicht. Hier sollte eine Vereinfachung stattfinden: Der DSH-Grammatiktest sollte entweder als eigenständiger Prüfungsteil aufgeführt werden, oder er sollte zusammen mit dem Leseverstehen einen gemeinsam bewerteten Prüfungsteil bilden.

Hinweise darauf, dass der DSH-Grammatiktest wesentlich zur Nützlichkeit der DSH beiträgt, haben sich in den Studien nicht ergeben. Nur wenn die Einheitlichkeit der Testmethoden-Merkmale gesichert wäre, könnte der DSH-Grammatiktest zur Erhöhung der Reliabilität beitragen. Durch einen Verzicht würde die Förderung der produktiven Grammatikkompetenz in der Vorbereitung voraussichtlich weniger Gewicht erhalten und das Testkonstrukt würde sich leicht verändern, aber die Zulassungsfunktion könnte auch ohne den DSH-Grammatiktest in ähnlicher Weise erfüllt werden. Würde man auf den DSH-Grammatiktest verzichten, wäre auch das Signal für die DSH als Feststellungsprüfung überzeugender und der Bezug zum prüfungsvorbereitenden Unterricht weniger stark.

5. Fachbezug in Sprachtests

Übersicht: Kapitel 5

Dem Thema Sprachtests mit Fachbezug sind die Kapitel 5 und 6 gewidmet. In Kapitel 5 stelle ich die Diskussion um Sprachtests mit Fachbezug vor; Kapitel 6 enthält eine eigene Studie zu diesem Thema. Am Anfang des Kapitels 5 steht die Frage, wie man in der DSH und im TestDaF mit dem Thema Fachbezug umgeht. Es folgt die Darstellung von Argumentationen, die für oder gegen den Einsatz von Sprachtests mit Fachbezug geführt werden (Kapitel 5.1). Forschungsergebnisse zur Rolle der Vorkenntnisse in Sprachtests mit Fachbezug stelle ich in Kapitel 5.2 vor.

Ausgangspunkt der Kapitel zum Fachbezug in Sprachtests für den Hochschulzugang sind wiederum Unterschiede zwischen dem TestDaF und der DSH. Im TestDaF verzichtet man ausdrücklich auf einen Fachbezug. Grotjahn beschreibt den Grundgedanken bei der Konzeption des TestDaF:

> Bei der Konstruktion der Verstehensaufgaben des TestDaF werden in erster Linie Aufgaben konstruiert, von denen angenommen wird, dass ihre Lösung lediglich gemeinsames Hintergrundwissen involviert (Grotjahn, 2000a: 17).

Die Umsetzung dieser Forderung dürfte auf Schwierigkeiten stoßen, da die Kandidaten, welche sich dem TestDaF unterziehen, aus der ganzen Welt kommen. Ob sich tatsächlich ein "gemeinsames Hintergrundwissen" ausmachen lässt, ist fragwürdig. Für den TestDaF erhebt Grotjahn daher die Forderung, den Einfluss der Variable "Hintergrundwissen" vorher prüfen zu lassen (Grotjahn, 2000a: 17). Derartige Erhebungen sind mir allerdings nicht bekannt.

Auch in der DSH soll ein Fachbezug vermieden werden. Im DSH-Handbuch heißt es zum Prüfungsteil Leseverstehen: "Der Text soll keine speziellen Fachkenntnisse voraussetzen" (FaDaF, 2001: 5/2). Die Rahmenordnung lässt einen Fachbezug unter bestimmten Umständen jedoch zu:

> Es soll ein weitgehend authentischer, studienbezogener und wissenschaftsorientierter Text vorgelegt werden, der keine Fachkenntnisse voraussetzt, ggf. nur solche, die Gegenstand eines vorangegangenen fachspezifisch orientierten Unterrichts waren (HRK/KMK, 2004: DSH-Musterprüfungsordnung, § 10).

Die Anweisung, Fachkenntnisse dürften im Leseverstehenstest nur dann eingesetzt werden, wenn sie Teil des prüfungsvorbereitenden Unterrichts seien, ist bemerkenswert. Die Aussage impliziert, dass ein Sprachtest mit Fachbezug Kandidaten mit Fachkenntnissen automatisch zu einem Vorteil verhilft bzw. Kandidaten ohne Fachkenntnisse benachteiligt. Die Aussage verdeutlicht, dass man davon ausgeht, dass eine DSH möglicherweise für eine bestimmte Zielgruppe erstellt wird. Die Zielgruppengenauigkeit ist einer der größten Vorteile der DSH gegenüber standardisierten und zentralen Prüfungen. Die Aussage ist schließlich ein deutlicher Hinweis darauf, dass die DSH durchaus die Funktion einer Kursabschlussprüfung annehmen kann.

Auch beim Thema Fachbezug in Sprachtests für den Hochschulzugang liegt der Ausgangspunkt für die Überlegungen also in einem Unterschied zwischen der DSH und dem TestDaF. Es liegen bereits eine Vielzahl an Studien zu Sprachtests mit Fachbezug vor. Positionen und Forschungsergebnisse zu diesem Thema stelle ich in Kapitel 5 vor. Kapitel 6 enthält eine eigene Studie zum Thema Sprachtests mit Fachbezug. In dieser Studie geht es um die Frage, ob der Einfluss der Vorkenntnisse vom Niveau der Fremdsprachenkenntnisse abhängt und ob sich in diesem Zusammenhang sprachliche Schwellen bestimmen lassen. Derartige Informationen könnten für die Testerstellung und –interpretation genutzt werden.

5.1. Begründungen und Problembereiche

Übersicht: Kapitel 5.1

In Kapitel 5.1 stelle ich die Diskussion um Sprachtests mit Fachbezug dar. Am Anfang stehen Aussagen zum Sprachgebrauch, Definitionen von Sprachtests mit Fachbezug sowie von Fachsprache(n). Es folgen drei Unterabschnitte zu Sprachtests ohne Fachbezug, Vorteilen von Sprachtests mit Fachbezug und zu Bedenken gegen Sprachtests mit Fachbezug.

Zum Sprachgebrauch: "Sprachtest mit Fachbezug" und "fachsprachlicher Test" halte ich für gleichbedeutend. Im Englischen wird der Ausdruck *specific purpose language test* verwendet. Was versteht man unter einem Sprachtest mit Fachbezug? Was unterscheidet einen Sprachtest mit Fachbezug von einem Test ohne Fachbezug? Nach Douglas (2000) spiegelt sich der Bezug zu einer fachlichen Sprachverwendungssituation in Inhalt und Vorgehen des Tests.

> A specific purpose language test is one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain (Douglas, 2000: 19).

Die Analyse der fachlichen Sprachverwendungssituation hat bei einem Sprachtest mit Fachbezug demnach nicht nur Auswirkungen auf die Auswahl authentischer Texte und Materialien. Auch die Aktivitäten, zu denen die Kandidaten durch den fachsprachlichen Test veranlasst werden, sollen realitätsnah und authentisch sein. Das Zusammenwirken von Sprachkompetenz und Fachkompetenz, welches bei Sprachtests ohne Fachbezug ausdrücklich vermieden werden soll, ist bei Sprachtests mit Fachbezug bewusst beabsichtigt. Ein weiterer, zentraler Unterschied zu Sprachtests ohne Fachbezug liegt in der Funktion von fachsprachlichen Tests: Sie sollen Informationen über die Fähigkeit der Kandidaten zum Umgang mit Sprache in bestimmten, fachlichen Sprach-

verwendungssituationen bieten. Die Aussagekraft von Sprachtests ohne einen expliziten Bezug zu Situationen, in denen die Sprache benutzt wird, ist mithin allgemeiner als von Sprachtests mit Fachbezug.

Die Definition von Douglas weist auch darauf hin, welches Testkonstrukt bei Sprachtests mit Fachbezug angenommen wird. Fachkenntnisse werden als Teil des Testkonstrukts angesehen, da die Kandidaten Fachkenntnisse ohnehin zur Bewältigung der fachsprachlichen Kommunikation außerhalb der Testsituation einsetzen. Sie können daher keine Störgröße sein, sondern sind Teil der Kompetenzen, zu denen der Test Auskunft geben soll. Dementsprechend äußert sich Douglas:

> Specific purpose language ability results from the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics (Douglas, 2000: 40).

Sprachtests mit Fachbezug basieren ebenso wie der Fachsprachenunterricht auf der Annahme, dass es unterscheidbare Sprachvarietäten gibt, die sich jeweils so stark unterscheiden, dass sich eine gesonderte Beschäftigung damit lohnt. Zur Bestimmung und Abgrenzung von Fachsprachen sind umfangreiche Untersuchungen durchgeführt worden. Hoffmann stellt mehrere Kategorien vor, nach denen Fachsprachen bestimmt werden können (2001: 533-537). Vom Standpunkt der Stilistik, welche den Zweck und die Wirkung von sprachlichen Äußerungen untersucht, wird der Fachstil über die Funktion einer Aussage abgegrenzt. Fachsprachen werden beispielsweise als "Stil der Wissenschaft" beschrieben. Gläser definiert den Fachstil dementsprechend als die "für die Gestaltung eines Fachtextes charakteristische Auswahl und Anordnung sprachlicher Mittel, die in einem Gesamtzusammenhang von Absicht, Inhalt, Form und Wirkung der Aussage fungieren" (Gläser, 1979: 26; zit. n. Hoffmann, 2001: 534). Zur Erfassung des Phänomens Fachsprache hält Hoffmann diesen Ansatz nicht für tragfähig, denn innerhalb des Fachstils werden keine weiteren Differenzierungen vorgenommen. Damit werde dieser Ansatz der inneren Differenziertheit von Fachsprachen nicht gerecht. Daneben existieren Bestimmungen von Fachsprachen als Varietät der Gesamtsprache. Von der Varietät einer Sprache spricht man, wenn sich gemeinsame Merkmale der verwendeten Sprache bestimmen lassen, ohne dass dabei völlig neue Teilsprachen entstehen. Voraussetzung für die Verwendung des Begriffs Varietät ist die Annahme eines bestimmten sprachlichen Standards, von dem sich Abweichungen beschreiben lassen. Eine varietätenorientierte Definition von Fachsprache lautet:

> Wir verstehen unter Fachsprachen heute die Variante der Gesamtsprache, die der Erkenntnis und begrifflichen Bestimmung fachspezifischer Gegenstände sowie der Verständigung über sie dient und damit den spezifischen kommunikativen Bedürfnissen im Fach allgemein Rechnung trägt (Möhn/Pelka, 1984: 26; zit. n. Hoffmann, 2001: 534).

Definiert man Fachsprache als "Subsprache", so wird die Unterordnung unter ein Ganzes betont. Bei diesem Ansatz wird der Kommunikationsgegenstand oder das Sachgebiet berücksichtigt. Auf diese Weise können weit reichende Differenzierungen vorgenommen werden. Die Fachsprachendefinition von Hoffmann, auf die man sich in der deutschsprachigen Fachliteratur häufig bezieht, basiert auf der Betrachtung von Fachsprachen als Subsprachen:

> Fachsprache – das ist die Gesamtheit aller sprachlichen Mittel, die in einem fachlich abgrenzbaren Kommunikationsbereich verwendet werden, um die Verständigung zwischen den in diesem Bereich tätigen Menschen zu gewährleisten (Hoffmann, 1985: 53).

Schließlich können Fachsprachen als Gruppensprachen differenziert werden, indem der Gebrauch der Sprache durch bestimmte Personengruppen als kennzeichnendes Merkmal hervorgehoben wird. Dieser Aspekt spielt auch bei der Differenzierung von Fachsprachen als Subsprache eine Rolle.

Zur Problematik von Sprachtests ohne Fachbezug

Sind Sprachtests ohne Fachbezug eine attraktive Alternative zu Sprachtests mit Fachbezug? Grundsätzlich ist diese Frage zu bejahen, obwohl sich aus der Ablehnung von Sprachtests mit Fachbezug (siehe Seite 193 ff) nicht zwangsläufig eine Argumentation für Sprachtests ohne Fachbezug ableiten lässt. Ich werde in diesem Abschnitt erläutern, dass der einfache Verzicht auf einen Bezug zu einem bestimmten Kommunikationsbereich meiner Ansicht nach nicht in jedem Fall überzeugt. Ich gehe dabei auf folgende problematische Aspekte ein:

- die Problematik, Sprachtests (für den Hochschulzugang) ohne Fachbezug zu entwickeln;
- die Problematik, einen studienvorbereitenden Sprachenunterricht ohne Fachbezug durchzuführen;
- die Problematik, wissenschaftssprachliche Eigenschaften unabhängig von einer Fachdisziplin zu bestimmen und zu vermitteln.

Zunächst möchte ich die Vorteile von Sprachtests ohne Fachbezug in Erinnerung rufen: Der wichtigste Pluspunkt liegt in einer vergleichsweise hohen Ökonomie. Weil die Tests so allgemein sind, ist ein spezieller Testzuschnitt auf bestimmte Zielgruppen nicht notwendig. Eine Testversion kann für alle Testteilnehmer erstellt werden. Bei Sprachtests für den Hochschulzugang findet ein Verzicht auf einen ausdrücklichen Fachbezug auch deshalb statt, weil man den Einfluss der Vorkenntnisse gering halten und so einen Beitrag zur Testfairness leisten möchte.

Zur Problematik, Sprachtests (für den Hochschulzugang) ohne Fachbezug zu entwickeln: Dies soll am Beispiel von Leseverstehenstests erläutert werden. Es wurde bereits darauf hingewiesen, dass es nicht einfach ist, Texte für Leseverstehenstests ohne Fachbezug zu finden, die nicht über einen Bezug zu irgendeiner Disziplin haben, deren Thema allen gleichermaßen bekannt bzw. unbekannt ist. Im DSH-Handbuch wird den Testerstellern empfohlen, einen "weitgehend authentische[n], studienbezogene[n] und wissenschaftsorientierte[n] Text zu wählen" (FaDaF, 2001: 5/2). Zum IELTS wird die Textauswahl wie folgt beschrieben:

> Texts are taken from magazines, journals, books, and newspapers. Texts have been written for a non-specialist audience. All the topics are of general interest. They deal with issues which are interesting, recognisably appropriate and accessible to candidates entering postgraduate or undergraduate courses (International English Language Testing System, 1999: 6).

Zur Textauswahl beim TestDaF erläutert die Projektgruppe TestDaF:

> Die Auswahl der Lesetexte orientiert sich an den zwei großen Anwendungsbereichen der Fremdsprache für die Adressaten: Hochschulalltag (soziales Umfeld) und allgemeiner wissenschaftlicher Lehr- und Lernbetrieb (Projektgruppe TestDaF, 2000: 68).

Alderson (1988) fragt, was mit dem Ausdruck "allgemeiner Text" eigentlich gemeint sei. Sind "allgemeine" Texte so neutral, dass ihr Thema allen unbekannt ist oder spielen die Vorkenntnisse doch eine Rolle? Gibt es auf dem geforderten Sprachniveau überhaupt Texte ohne Fachbezug, Texte, welche nicht in irgendeiner Weise an ein Fach gebunden sind? In der Praxis gibt es zwei Reaktionen auf diese Fragen: Zunächst scheint man häufig auf Texte aus den Geisteswissenschaften zurückzugreifen, wohl weil man bei geisteswissenschaftlichen Themen davon ausgeht, dass sie allen Kandidaten zugänglicher sind. Alderson vermutet, dass Testersteller – häufig Lehrkräfte im Sprachenbereich – Texte aus den Geisteswissenschaften aufgrund ihrer eigenen Nähe zu geisteswissenschaftlichen Inhalten als neutral einschätzen (Alderson, 1988). Abgesehen von Texten, die im weitesten Sinne mit Computertechnik oder Ökologie zu tun haben,

scheinen Texte aus technischen oder naturwissenschaftlichen Disziplinen in Leseverstehenstests "ohne Fachbezug" in der Tat selten zu sein. Eine weitere Reaktion auf das Problem der fachlichen Neutralität ist der Einsatz mehrerer Texte. Im TOEFL, IELTS oder TestDaF werden in Prüfungsteilen zum Leseverstehen jeweils mehrere Texte mit unterschiedlicher inhaltlicher Ausrichtung eingesetzt. Auf diese Weise wird eine inhaltliche Tendenz verringert. Es führt jedoch auch dazu, dass sehr kurze Texte eingesetzt werden, an denen die Fähigkeit, mit längeren Texten umzugehen, nur indirekt geprüft werden kann.

Es soll nicht der Eindruck erweckt werden, als sei die Erstellung von Sprachtests ohne Fachbezug ein unsinniges Unterfangen. Es sei aber darauf hingewiesen, dass auch bei der Konstruktion "fachneutraler" Sprachtests schwierige Klippen umschifft werden müssen. Beachtenswert sind außerdem mögliche Auswirkungen von Sprachtests für den Hochschulzugang ohne Fachbezug. Man verzichtet auf das Signal, das ein Sprachtest mit Fachbezug auf die Testvorbereitung ausübt. Ein Signal, das möglicherweise zu einer verstärkten Berücksichtigung von fachsprachlichen Elementen im studienvorbereitenden Sprachenunterricht führt.

Zur Problematik, einen studienvorbereitenden Sprachenunterricht ohne Fachbezug durchzuführen: Die Diskussion um den studienvorbereitenden Sprachenunterricht umfasst mehrere Aspekte. Zunächst geht es um Zuständigkeiten: In der englischsprachigen Sprach- und Unterrichtswissenschaft wird "Englisch für allgemeine Studienzwecke" (*English for General Academic Purposes*) inhaltlich mit der Vermittlung von Lern- und Studiertechniken gleichgesetzt (Jordan, 1997; Dudley-Evans/St John, 1998). Die Vermittlung von Lern- und Studiertechniken spielt in der Studienvorbereitung von ausländischen Studienbewerbern fraglos eine wichtige Rolle. Ich bin jedoch der Ansicht, dass ein Unterricht, der in keiner Weise zwischen verschiedenen Fachdisziplinen differenziert, nur eine Hilfskonstruktion darstellt; eine Hilfskonstruktion, welche eine Antwort auf die organisatorischen Schwierigkeiten einer Differenzierung nach Fachdisziplinen, jedoch keine adäquate Vorbereitung auf die sprachlichen Anforderungen des Studiums darstellt. Auch die Vermittlung von Lern- und Studiertechniken sollte an Inhalte geknüpft sein. Strategien zum Wortschatzerwerb sollten nicht nur exemplarisch vorgeführt, sondern auch an konkreten Inhalten exerziert werden. Lesestrategien sollten nicht an beliebigen Texten aus dem Studienumfeld,

sondern an authentischen Texten aus dem Studium geübt werden. Gleiches gilt für die Planung von Referaten, das Verfassen von Texten, das Mitschreiben von Vorlesungen oder die Beteiligung an Diskussionen. Die Anbindung an eine Fachdisziplin ist für die Teilnehmer nicht nur motivierender, sie ist auch notwendig, weil sich die Anforderungen von Fach zu Fach unterscheiden. In den Geisteswissenschaften gehört die Versuchsbeschreibung beispielsweise nicht zu den sprachlichen Aufgaben des Studiums, in den Naturwissenschaften schon. Die Vermittlung von Lern- und Arbeitstechniken sollte daher auch im studienvorbereitenden Sprachenunterricht in die Vermittlung der jeweiligen Fachsprache integriert werden.

Zur Problematik, wissenschaftssprachliche Eigenschaften unabhängig von einer Fachdisziplin zu bestimmen und zu vermitteln: Ein weiterer Aspekt der Diskussion um studienvorbereitenden Sprachenunterricht bezieht sich auf den Gegenstand. In Deutschland gibt es das Bemühen, wissenschaftssprachliche Eigenschaften unabhängig von einer Fachdisziplin zu bestimmen und zu vermitteln (Ehlich, 1994; 1999; 2000; Graefen, 1997; 2000). Überlegungen zur "allgemeinen Wissenschaftssprache" beruhen auf zwei Anliegen: Zum einen ist die Förderung des Deutschen als internationale Sprache der Wissenschaften zu nennen. Ehlich (1999; 2000) bemängelt etwa, dass Deutsch in der internationalen Wissenschaftskommunikation keine Rolle (mehr) spielt. Zum anderen beruhen die Überlegungen zur allgemeinen Wissenschaftssprache auf den praktischen Notwendigkeiten des studienvorbereitenden und –begleitenden Sprachenunterrichts. Es gibt verschiedene Möglichkeiten, die Lerngruppen zusammenzustellen: Man kann sich nach Merkmalen der Teilnehmer orientieren und das Niveau der Deutschkenntnisse als Grundlage nehmen, man kann das (angestrebte) Studienfach oder sogar das Herkunftsland als Grundlage für eine Gruppenzusammenstellung nehmen. Häufig sind Kurse zu bestimmten Fertigkeiten. Wenn man sich die Programme einiger universitärer Sprachlehrgebiete ansieht, stellt man fest, dass Gruppen nach dem Niveau der Deutschkenntnisse, nach bestimmten Fertigkeiten oder auch nach Prüfungsziel (DSH, TestDaF) zusammengestellt werden. Eine Differenzierung nach dem zukünftigen Studienfach wird außerhalb von Studienkollegs selten vorgenommen. Angesichts organisatorischer Schwierigkeiten wäre es also durchaus praktisch, wenn sich Merkmale des universitären bzw. wissenschaftlichen Sprachgebrauchs unabhängig von einem bestimmten Fach, von bestimmten Inhalten vermitteln ließen.

Ist die "Wissenschaftssprache" eine abgrenzbare Varietät der Gesamtsprache oder gar eine eigene Fachsprache? Unterschiedliche Kommunikationsformen in unterschiedlichen Fächern führen meiner Ansicht nach dazu, dass es nicht produktiv ist, sich im studienvorbereitenden Sprachunterricht und in Sprachtests für den Hochschulzugang auf Phänomene der "allgemeinen Wissenschaftssprache" zu konzentrieren. Das Phänomen der Wissenschaftssprache findet in der Fachsprachenforschung gleichwohl einige Beachtung. Wissenschaftssprache wird in der Regel als Oberbegriff für verschiedene Fachsprachen interpretiert (Kalverkämper, 1998; Kretzenbacher, 1992; 1998). Es gibt mehrere Studien, in denen Fachkommunikation mit dem Ziel analysiert wurde, fächerübergreifende sprachliche Phänomene zu beschreiben (z. B. Beneš, 1971; Erk, 1972; 1975; 1982; 1985). Die Bezeichnungen "alltägliche Wissenschaftssprache" bzw. "wissenschaftliche Alltagssprache" gehen auf einen Vorschlag von Ehlich (1994) zurück. Die Betrachtung der Wissenschaftssprache ohne Fachbezug beruht auf verschiedenen Anliegen. Sie ist einerseits eine Reaktion auf die häufige Konzentration der Fachsprachenforschung und des Fachsprachenunterrichts auf die Fachterminologie. Es ist unbestritten, dass es eine unzureichende Beschreibung und Analyse fachsprachlicher Phänomene ist, wenn der Fachfremdsprachenunterricht mit der Vermittlung von Terminologie gleichgesetzt wird. Ehlich fordert folgerichtig eine Konzentration auf andere Phänomene, welche ebenfalls wesentlich zur erfolgreichen Kommunikation in den Wissenschaften beitragen. Sein Vorhaben beschränkt sich jedoch nicht auf die Kommunikationsverfahren innerhalb einer bestimmten Disziplin. Er glaubt vielmehr für die Sprache(n) der Wissenschaften gemeinsame Merkmale identifizieren zu können, welche eine eigene Betrachtung verdienen.

> Die Verwendungsmöglichkeit von Ausdrücken wie "einen Grundsatz ableiten" oder "eine Erkenntnis setzt sich durch" macht den Wissenschaftler mindestens ebenso aus wie die genaue Kenntnis seiner eigenen Fachterminologie. Anders formuliert: die Mitgliedschaft, die zur wissenschaftlichen Kommunikation befähigt und ermächtigt, ergibt sich gerade auch über die passive und aktive Beherrschung dieser Facetten von Wissenschaftssprache, die auf den ersten Blick als allgemeinsprachliche erscheinen, es in Wahrheit aber nicht sind (Ehlich, 1994: 339-340).

Graefen weist in ihrer Definition auf das Grundproblem hin: Es ist die Häufigkeit, mit der Wendungen und Strukturen auftreten. Ein studienvorbereitender Sprachenunterricht muss sich auf häufige und relevante Phänomene konzentrieren.

> Die Bezeichnung 'Alltägliche Wissenschaftssprache' erfaßt, grob gesagt, denjenigen Anteil der für wissenschaftliche Zwecke verwendeten Sprache, der in allen Fächern bekannt, verwendbar und – mehr oder weniger frequent – auch in Gebrauch ist (Graefen, 2000: 191).

Ich fasse zusammen: Die Vermittlung einer "allgemeinen Wissenschaftssprache" ohne Bezug zu den unterschiedlichen sprachlichen Konventionen und Anforderungen fachlicher Subsprachen hat organisatorische Vorteile, da die Gruppeneinteilung erleichtert wird. Es ist aber eine Hilfskonstruktion, welche aus organisatorischen Notwendigkeiten, einer Ablehnung von überzogener Terminologieorientierung des Fachfremdsprachenunterrichts und der Sorge um die Stellung der deutschen Sprache entstanden ist, nicht aber aus einer überzeugenden Analyse des Sprachbedarfs ausländischer Studierender.

In Abwesenheit einer plausibleren Konzeption für einen Unterricht in der allgemeinen Wissenschaftssprache bleiben auch zentrale Fragen an Sprachtests für den Hochschulzugang, welche ohne einen Fachbezug auskommen, unbeantwortet: Welche Varietät der Wissenschaftssprache ist Gegenstand der Prüfung? Und mit welcher Subsprache sollen sich ausländische Studienbewerber in der Prüfungsvorbereitung auseinander setzen? Welche Sprache soll im studien- und prüfungsvorbereitenden Sprachunterricht vermittelt werden? Ob Sprachtests ohne Fachbezug, die ohne Unterschied auf das gewünschte Studienfach für alle Kandidaten gleich sind, ein aussagekräftigerer Prädiktor für sprachliche Leistungen im Fachstudium darstellen, ist nicht gewiss. Aus der Argumentation gegen Sprachtests für den Hochschulzugang *mit* Fachbezug (siehe Seite 193 ff) lässt sich meiner Ansicht nach kaum eine Unterstützung für Sprachtests *ohne* Fachbezug ableiten.

Ich schließe das Kapitel mit einem Hinweis auf Unterrichtskonzeptionen zum universitären Fachfremdsprachenunterricht. Diese sind zwar nicht unumstritten, sie stützen sich jedoch in der Regel auf eine solide Analyse des Sprachbedarfs und stoßen meiner Einschätzung nach bei ausländischen Studienbewerbern und ausländischen Studierenden auf eine hohe Akzeptanz (für den deutschsprachigen Bereich z. B. Buhlmann/Fearns, 2000; Fluck, 1992, 1996; Hoffmann/Kalverkämper/Wiegand, 1998; 1999; Monteiro, 1990; Schröder, 1988). Als Ziel des Fachsprachenunterrichts bestimmen Buhlmann und Fearns die "Sprachliche Handlungsfähigkeit im Fach", die an Fachinhalten geschult werden soll (Buhlmann/Fearns, 2000: 9). Der Fachfremdsprachenunterricht übernimmt dabei eine Brückenfunktion zwischen dem Fremdsprachenunterricht und dem Fachunterricht. Dies halte ich für eine sinnvolle Vorgehensweise.

Vorteile von Sprachtests mit Fachbezug

Welche Vorteile, welchen Nutzen verspricht man sich vom Einsatz von Sprachtests mit Fachbezug? In diesem Abschnitt stelle ich mehrere Argumentationen vor:

- zum Testkonstrukt und zur Validität;
- zum Testkonstrukt und zur Testfairness;
- zur Authentizität/Direktheit von Sprachtests mit Fachbezug;
- zur Testerstellung und Reliabilität;
- zu Testauswirkungen auf die Testvorbereitung;
- zur Problematik von Sprachtests ohne Fachbezug.

Bei diesen Argumentationen kommt es zu inhaltlichen Überschneidungen.

Testkonstrukt und Validität: Nach Douglas kann vor allem die Validität für den Einsatz von Sprachtests mit Fachbezug sprechen: Experten verwenden zur Kommunikation in ihrem Fach eine Sprache mit besonderen Eigenschaften. Wer in diesem Fach tätig ist, muss mit der Fachsprache umgehen können (Douglas, 2000: 7-8). Mit Blick auf Sprachtests für den Hochschulzugang bedeutet dies: Studierende müssen mit der Fachsprache in ihrem Fach umgehen können. Dass dies jedoch nicht vor Aufnahme des Fachstudiums geprüft werden soll, machen Kritiker geltend (s. u.). Douglas geht weiter davon aus, dass die sprachliche Leistung in Abhängigkeit vom Kontext und vom Aufgabentyp variiert. Das Urteil über die Sprachkompetenz wird demnach ebenfalls unterschiedlich ausfallen. Wenn es das Ziel des Testverfahrens ist, Aussagen zu treffen über den Grad der Sprachbeherrschung in einem fachlichen Zusammenhang, dann ist der Einsatz eines realitätsnahen Sprachtests mit Fachbezug sinnvoll. Diese Argumentation gilt vor allem für Sprachtests für den Beruf. Aber auch bei Sprachtests für den Hochschulzugang dürfte ein Test, der Sprachverwendungssituationen aus dem Fachstudium aufgreift, einen besseren Indikator für Sprachkompetenz im Studium darstellen als Tests, die darauf verzichten.

Testkonstrukt und Testfairness: Im Sinne der Argumentation von Douglas erhöht der Einsatz von Sprachtests mit Fachbezug unter bestimmten Umständen die Fairness. Wenn tatsächlich der Sprachgebrauch in einem Fach geprüft werden soll, sollten Kandi-

daten, welche über eine hohe fachsprachliche Kompetenz verfügen, diese im Sprachtest auch einsetzen können. So argumentieren auch Alderson und Urquhart. Sie verbinden die Schlussfolgerung mit der Annahme, dass Kandidaten mit guten Fachsprachenkenntnissen ihre wahre Kommunikationsfähigkeit im Fach in Sprachtests ohne Fachbezug nicht unter Beweis stellen können:

> Should it be found, however, that general tests were discriminating against a major group, say engineers, or that they were having the effect of denying further study to students who were quite competent readers in their own academic area, then these practical advantages would not be enough to ensure the survival, in tertiary ESP, of general tests (Alderson/Urquhart, 1985b: 27-28).

Ob Sprachtests mit Fachbezug besonders fair sind, hängt von der Interpretation des Testkonstrukts ab: Sollen nur Sprachkenntnisse erhoben werden, oder wird die sprachliche Handlungsfähigkeit im Fachgebiet als Testkonstrukt angesehen? Wenn Sprachtests so konstruiert sind, dass Kandidaten mit Vorkenntnissen ein besseres Ergebnis erzielen, spielt das Testkonstrukt die entscheidende Rolle für die Legitimation. Wenn sich ein Test auf Sprache in einem beruflichen oder wissenschaftlichen Fachgebiet bezieht, ergibt sich die zu messende Fertigkeit – der gekonnte Einsatz der Sprache – aus einem Zusammenspiel von Sprach- und Fachkompetenzen. Denkbar ist, dass Vorkenntnisse nicht als Störgröße angesehen werden, welche das Konstrukt verfälschen, sondern als unumgänglicher Teil des Konstrukts anerkannt werden. Es ist – so lautet etwa die Argumentation von Alderson (oder Douglas, s. o.) – notwendig, die Rolle der Vorkenntnisse zur Kenntnis zu nehmen und sie bei der Testkonstruktion zu berücksichtigen. Sie können einen Beitrag dazu leisten, dass vorhandene sprachliche Leistungen überhaupt abgerufen werden. Zum Leseverstehen formuliert er:

> Background knowledge should be recognised as influencing all comprehension, and therefore every attempt should be made to allow background knowledge to facilitate performance, rather than allowing its absence to inhibit performance (Alderson, 2000a: 29).

Authentizität und Direktheit: Bei Sprachtests mit Fachbezug leitet man Inhalt und Aufgabenstellungen ab aus einer Analyse der sprachlichen Anforderungen, denen eine Person in einer bestimmten Situation gewachsen sein muss. Realitätsnahe Sprachtests, die als Performanztests angelegt sind, lassen eine hohe Augenscheinvalidität erwarten (Douglas, 2001: 172). Auch Alderson vermutet, dass Texte und Themen aus dem Studienfach auf ein größeres Interesse treffen und damit auch eine intensivere Auseinandersetzung auslösen (Alderson, 2000a: 29). Die Direktheit von Sprachtests mit Fachbezug führt auch zu der Erwartung einer hohen Konstruktvalidität. Die Argumen-

tation wurde in Kapitel 2.1 (Seite 29) im Zusammenhang mit der Klassifikation von Sprachtests bereits vorgestellt.

Testerstellung und Reliabilität: Eine weitere positive Annahme bezieht sich auf die Erstellung von Sprachtests mit Fachbezug. Bei Leseverstehenstests wird häufig eine geringe Paralleltestreliabilität beobachtet. Alderson stellte in einer Studie mit Leseverstehenstests und Grammatiktests beispielsweise zu seiner Überraschung fest, dass die Ergebnisse in den Grammatiktests höher mit den Leseverstehenstests korrelierten als die Leseverstehenstests untereinander (1993). Äquivalente Formen von Leseverstehenstests kommen selten vor, weil das Testkonstrukt von Leseverstehenstests durch mehrere Faktoren bestimmt wird. Die Verständlichkeit der Texte und die unterschiedlichen Items tragen maßgeblich zum Schwierigkeitsgrad eines Tests bei. Eine Einschätzung der Verständlichkeit unterschiedlicher Texte ist jedoch schwierig. Möglicherweise treten die Schwierigkeiten bei der Erstellung von Leseverstehenstests mit Fachbezug nicht in gleichem Maße auf. Da sich Leseverstehenstests mit Fachbezug auf (mehr oder weniger) abgrenzbare Kommunikationssituationen beziehen, dürfte es leichter sein, passende Testvorgaben zu finden. Unter diesen Umständen ist die Konstruktion von äquivalenten Leseverstehenstests vermutlich eher möglich.

Testauswirkungen auf die Testvorbereitung: Ein weiterer zentraler Aspekt, auf den hinzuweisen ist, ist der Zusammenhang zwischen Sprach*tests* mit Fachbezug und dem Fachsprachen*unterricht*. Die Entwicklung von Fachsprachentests hängt eng mit dem Aufkommen des Fachfremdsprachenunterrichts zusammen. Ursprünglich sind Sprachtests mit Fachbezug im Fahrwasser des Fachfremdsprachenunterrichts entwickelt worden. Bei der Auswahl der Inhalte und Aufgaben von fachsprachlichen Tests wurde nach ähnlichen Prinzipien wie beim Fachsprachenunterricht vorgegangen. Mit dem Einsatz einer (eher) direkten Vorgehensweise bei Sprachtests mit Fachbezug hängt auch die Erwartung zusammen, dass ein Sprachtest mit Fachbezug positive Auswirkungen auf die Lehr- und Lernprozesse ausübt. Sprachtests mit Fachbezug können als Signal für die Prüfungsvorbereitung verstanden werden. Sie legen eine sprachliche Studienvorbereitung nahe, die auch Fachsprache einbezieht. Ich halte eine Studienvorbereitung, welche Fachinhalte und Sprachvermittlung verbindet, aus mehreren Gründen für sinnvoll: Sprachunterricht in der Studienvorbereitung ist Unterricht mit fortgeschrittenen Lernern, Sprachunterricht in der Studienvorbereitung ist Unterricht für Erwachsene; ein

Bezug zum (angestrebten) Studienfach wird von den Teilnehmern in der Regel gut geheißen und begrüßt. Auch inhaltlich ist die Förderung der Sprachkompetenz an Fachthemen in Verbindung mit Fachinhalten folgerichtig.

Zur Problematik von Sprachtests ohne Fachbezug: Ein Argument für den Einsatz von Sprachtests mit Fachbezug lässt sich schließlich aus Mängeln ableiten, die mit möglichen Alternative zusammenhängen: Wie nützlich sind Sprachtests ohne Fachbezug für den Hochschulzugang? Sind Sprachtests ohne Fachbezug für weit fortgeschrittene Lerner überhaupt möglich? Dieses Argument stellte ich bereits im Abschnitt "Sprachtests ohne Fachbezug" vor (Seite 183 ff).

Bedenken gegen Sprachtests mit Fachbezug

Die Argumente für fachsprachliche Tests haben sich durchgesetzt, allerdings nur bei Sprachtests für berufliche Zwecke. Bei den meisten Sprachtests für das Studium verzichtet man auf einen Fachbezug. Folgende Argumente werden angeführt:

- Ablehnung des Fachfremdsprachenunterrichts;
- Problematik der Differenzierung zwischen Fachsprachen;
- Problematik der Testökonomie;
- Problematik der Direktheit von Sprachtests;
- Problematik der Testfairness;
- Problematik von Fachkenntnissen in Sprachtests für den Hochschulzugang.

Ablehnung des Fachfremdsprachenunterrichts: Skehans Ablehnung von Sprachtests mit Fachbezug basiert auf einer Ablehnung des Fachfremdsprachenunterrichts (1984). Er bemängelt, dass der Fachsprachenunterricht sich häufig an strukturellen Phänomenen orientiere und auf einer "atomistischen" Sicht von Sprache beruhe. Wie auch andere Autoren vermisst er eine empirische Basis für das Projekt des Fachfremdsprachenunterrichts im Allgemeinen und das fachsprachliche Testen im Besonderen.

Problematik der Differenzierung zwischen Fachsprachen: Die ebenfalls grundsätzliche Kritik von Davies (2001) richtet sich gegen die Annahme, dass man unter den Fachsprachen Differenzierungen vornehmen kann, die für Sprachtests von Belang sind. Man muss in der Tat zur Kenntnis nehmen, dass Fachsprachen sich zwar zusammenfassen und beschreiben lassen (je nach Blickrichtung als Varietät, Subsprache oder Gruppensprache, s. o.), dass sie sich im konkreten Fall aber nicht eindeutig abgrenzen lassen. Hier hilft man sich mit der Vorstellung eines Kontinuums mit unterschiedlichen Ausprägungen, dem sich einzelne Texte zuordnen lassen.

Bei Sprachtests für den Hochschulzugang ist eine Differenzierung von Fachsprachen als Subsprachen von besonderem Interesse, da sie wichtige Fragen aufwirft: Wie abgrenzbar ist der Kommunikationsbereich "Sprachverwendung während des Studiums des Faches XY"? Kann die Gesamtheit aller sprachlichen Mittel für die Kommunikationssituation Hochschulstudium angemessen beschrieben werden? Lässt

sich aus dieser Beschreibung eine repräsentative Auswahl an sprachlichen Mitteln für den Sprachtest ableiten? Anders als im beruflichen Alltag, in dem sich bestimmte Kommunikationsformen häufig präzise beschreiben lassen, sind Kommunikationsformen im Studium weniger gut fassbar. Das Spektrum der Sprachverwendungssituationen ist breit. Hinzu kommt, dass ein Studium nicht nur aus einem Thema besteht, sondern vielfältige Inhalte betrifft.

Problematik der Testökonomie: Die große Anzahl verschiedener Fächer und die schwierige Abgrenzung zwischen einzelnen Disziplinen führen zu Schwierigkeiten in der Durchführung. Daher bestehen auch praktische und ökonomische Bedenken gegen den Einsatz von Sprachtests mit Fachbezug. Ist es wirklich sinnvoll und ökonomisch, für verschiedene Studienfächer unterschiedliche Sprachtests zu entwerfen? Alderson und Urquhart geben zu bedenken:

> They [specific purpose language tests] are inevitably more expensive and more difficult to administer. The number of specialist modules is very debatable: do we have a test for all engineers or one for chemical engineers, one for electronic engineers, etc? (Alderson/Urquhart; 1985b: 27).

Allerdings stellen Alderson und Urquhart auch fest, dass sie die vier unterschiedlichen Fachbezüge eigentlich nicht gebraucht hätten, dass drei oder sogar nur zwei Versionen ausgereicht hätten (siehe Kapitel 5.2).

Problematik der Direktheit von Sprachtests: Vorbehalte gibt es auch gegen die mit dem Einsatz von direkten Sprachtests verbundene Erwartung, diese würden das Problem der Konstruktvalidität lösen (Bachman, 1990; Clapham, 2000; Fulcher, 1999). Auch der Einsatz von vermeintlich direkten Sprachtests entbindet in der Tat nicht davon, weitere Argumente zur Validität eines Sprachtests zu suchen.

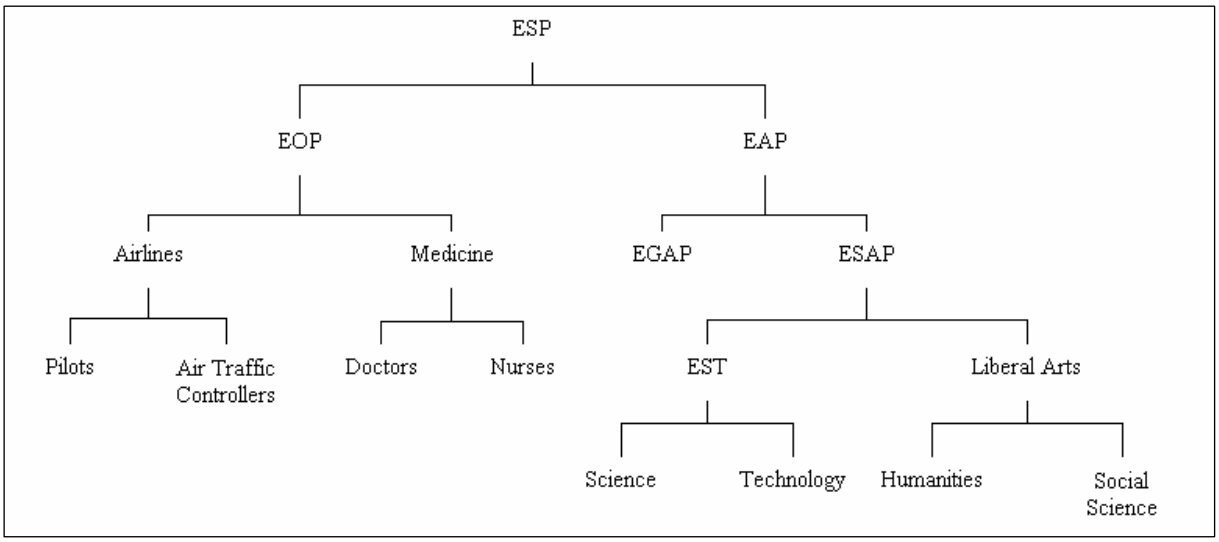
Problematik der Testfairness: Weitere Autoren beziehen sich auf den Einfluss der Fachkenntnisse auf die Leistungen in Sprachtests mit Fachbezug. Wenn Sprachtests Fachkenntnisse voraussetzen, sind Kandidaten ohne derartige Vorkenntnisse möglicherweise gegenüber Kandidaten mit Vorkenntnissen benachteiligt. Vorbehalte gegen den Einsatz von Sprachtests mit Fachbezug gibt es, weil befürchtet wird, dass die Berücksichtigung des Faches und die Rolle der Fachkenntnisse auch die Ergebnisse verändern.

Wenn Fachkenntnisse einen Einfluss auf das Ergebnis von Sprachtests haben, dann werden nicht nur Sprachkenntnisse, sondern auch Fachkenntnisse geprüft. Dies ist nicht

fair, wenn die Kandidaten noch nicht über Fachkenntnisse verfügen. Der Einfluss der Vorkenntnisse ist das zentrale Argument gegen den Einsatz von Sprachtests mit Fachbezug. Häufig wird die Ansicht vertreten, dass Fachkenntnisse nicht zum Testkonstrukt von Sprachtests gehören, auch nicht von Sprachtests mit Fachbezug:

> LSP [Language for Specific Purposes] testing cannot be about testing for subject specific knowledge (Davies, 2001: 143).
>
> ...background knowledge must be controlled so that it will not account for an indeterminate amount of assessment results [...] to factor reader background out of assessment [...] is to assess something other than reading comprehension (Farr/Carey/Tone, 1985: 140).

Vorkenntnisse werden als Störgröße angesehen, welche möglichst auszuschließen ist. Sie werden nicht als unumgänglichen Teil des Testkonstrukts angesehen, mit dessen Hilfe das Zustandekommen einer sprachlichen Leistung unterstützt wird.

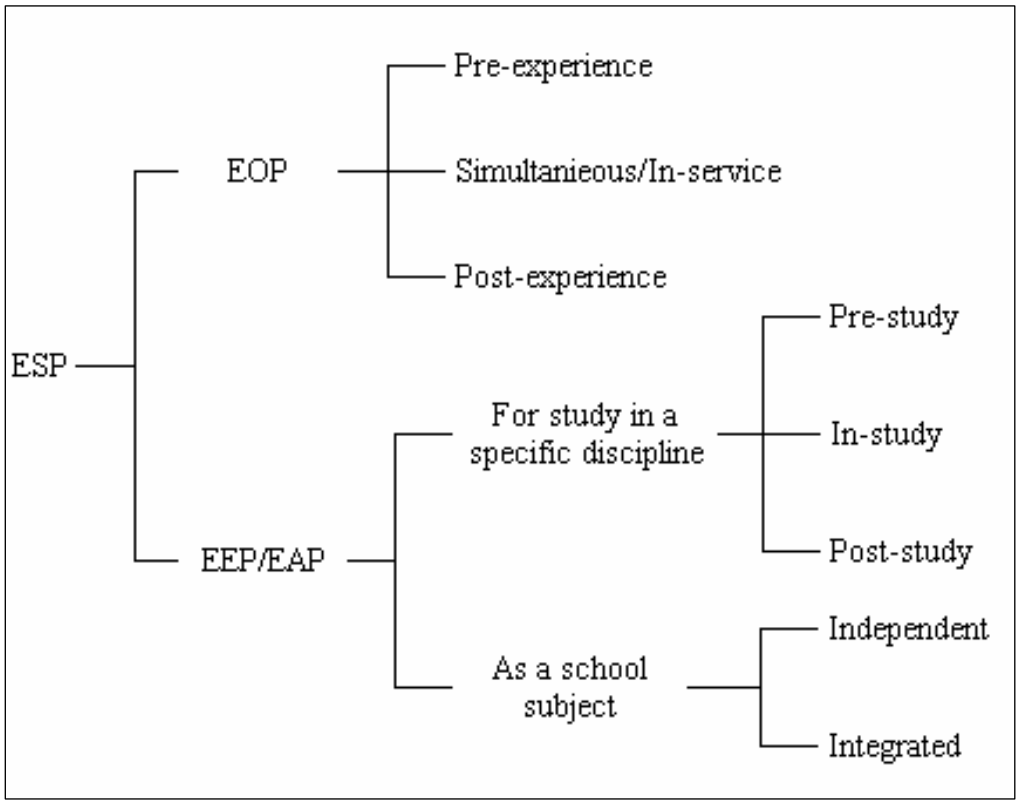


Quelle: Jordan, 1997: 3.

Abbildung 18: Klassifizierung der Fachsprache Englisch nach Subsprachen

Problematik von Fachkenntnissen in Sprachtests für den Hochschulzugang:
Argumente gegen den Einsatz von Sprachtests mit Fachbezug gründen auf dem Zeitpunkt, zu dem der Test durchgeführt wird. Sprachtests für den Hochschulzugang finden eben vor der Aufnahme des Studiums statt, daher können Fachkenntnisse noch nicht vorausgesetzt werden. Diese Abgrenzung lässt sich anhand der Abbildung 18 erläutern, welche eine Klassifizierung von Fachsprachen als Subsprachen am Beispiel von Englisch zeigt.

Jordan unterteilt Fachenglisch nach der Funktion zunächst in Englisch für den Beruf (*English for Occupational Purposes, EOP*) und Englisch für das Studium bzw. die Wissenschaften (*English for Academic Purposes, EAP*), (Jordan, 1987; auch Dudley-Evans/St John, 1998: 6; siehe Abbildung 18). Bei Englisch für den Beruf unterscheidet er zwischen Arbeitsbereichen (z. B. Luftfahrt oder Medizin) und weiter nach Berufsgruppen (z. B. Piloten und Fluglotsen oder Ärzte und Krankenpfleger). Weitere Differenzierungen sind vorstellbar. Englisch für das Studium werden in Englisch als allgemeine Wissenschaftssprache bzw. Englisch für allgemeine Studienzwecke (*English for General Academic Purposes, EGAP*) und Englisch als spezielle Wissenschaftssprache bzw. Fachenglisch (*English for Specific Academic Purposes, ESAP*) getrennt. Auch bei den Fächern lassen sich Unterscheidungen vornehmen, als Oberbegriff etwa Englisch für Naturwissenschaften und Technologie (*English for Science and Technology, EST*) sowie weiter Englisch für Naturwissenschaften und Englisch für Technologie. Auch hier sind weitere Differenzierungen nach Studienfächern und sogar einzelnen Studieninhalten denkbar: Fachenglisch für Chemie, Fachenglisch für Biologie. Und weiter: Fachenglisch für Ökologie und Fachenglisch für Mikrobiologie. Diese Subsprachen lassen sich beschreiben und mehr oder weniger genau voneinander abgrenzen.



Quelle: Robinson, 1991: 3-4.

Abbildung 19: Klassifizierung des Fachsprachenunterrichts (Englisch)

Die Ablehnung eines Fachbezugs in Sprachtests für den Hochschulzugang mit dem Argument des Testzeitpunkts wird auch durch das Schaubild von Robinson (1991) verdeutlicht (siehe Abbildung 19, Seite 197). Er klassifiziert den Fachsprachen-*unterricht* Englisch wiederum in Englisch für den Beruf (*English for Occupational Purposes, EOP*) und Englisch für das Studium bzw. die Ausbildung (*English for Academic Purposes, EAP; English for Educational Purposes, EEP*). Interessant ist bei Sprachtests für den Hochschulzugang die Unterscheidung zwischen Sprachenunterricht vor der Aufnahme des Studiums, während des Studiums und im Anschluss an das Studium.

Zum Abschluss des Kapitels 5.1 ("Begründungen und Problembereiche") möchte ich die Argumente herausgreifen, die meiner Ansicht nach zentral sind: Ich halte die Auswirkungen auf die Testvorbereitung und die Authentizität für die wichtigsten Vorteile

von Sprachtests mit Fachbezug. Ich bin außerdem der Ansicht, dass Sprachtests ohne Fachbezug keine überzeugende Alternative darstellen. Die Argumentation für fachsprachliche Tests überzeugt mich vor allem, wenn es um Sprachtests mit Bezug zu beruflichen Kommunikationssituationen geht. Sprachtests für Berufssprachen können sich auf eine Subsprache beziehen, die sich von anderen abgrenzen lässt. Englisch für Fluglotsen oder Deutsch für die Börse sind derartige Subsprachen. Wenn es um fachsprachliche Tests für den Hochschulzugang geht, sind Bedenken angebracht. Problematisch ist der Gegenstand der Tests: Lassen sich Sprachen im Studium überhaupt ausreichend differenzieren? Unklar ist auch die Rolle der Vorkenntnisse: Warum soll man Studienbewerbern mit Kommunikationssituationen aus dem Studium konfrontieren? Möglicherweise ist die Testfairness eingeschränkt. Überhaupt stehen die Fragen nach dem Testkonstrukt und der Fairness im Mittelpunkt der Diskussion. Während Gegner von fachsprachlichen Tests Vorkenntnisse grundsätzlich nicht als Teil des Testkonstrukts ansehen und es vermeiden möchten, dass Testteilnehmer Vorkenntnisse zum Thema einbringen können, gehen Befürworter davon aus, dass Vorkenntnisse eine Kommunikation im Fach erst ermöglichen und dazu führen, dass die Testteilnehmer ihre besten Leistungen zeigen können. Eine vermittelnde Position formulieren Urquhart und Weir:

> One of the main causes of differing interpretations is background knowledge, and the elimination of this variable would seem an obvious step. For most practical purposes, however, this is likely to be an impossibility; the theory would suggest that background knowledge is always present. All we can do is attempt to minimise the effect of that variable (Urquhart/Weir, 1998: 115).

Ich möchte den Vorschlag von Urquhart und Weir aufgreifen – allerdings in leicht veränderter Form: Es geht mir nicht darum, die Variable Vorkenntnisse zu minimieren, sondern besser zu verstehen, welchen Einfluss Vorkenntnisse auf Leistungen in fachsprachlichen Tests haben. Dazu stelle ich Forschungsergebnisse (Kapitel 5.2) und ein eigenes Forschungsprojekt vor (Kapitel 6).

5.2. Forschungsergebnisse

Übersicht: Kapitel 5.2
Kapitel 5.2 enthält eine Übersicht über Forschungsergebnisse zu Sprachtests mit Fachbezug. Im Mittelpunkt stehen zwei Fragen: Erstens, welche Rolle die Vorkenntnisse bei Sprachtests mit Fachbezug spielen, zweitens, welche anderen Variablen die Intensität der Einflussvariable Vorkenntnisse beeinflussen. Hier geht es vor allem um die "Doppelte Schwellenhypothese", nach der die Rolle von Vorkenntnissen von dem Niveau der Fremdsprachenkenntnisse abhängt. Da sich Studien zur Rolle der Vorkenntnisse in Sprachtests mit Fachbezug häufig mit Tests zum Leseverstehen beschäftigen, beginnt das Kapitel mit Aussagen zum Lesen.

Die meisten Sprachtests für den Hochschulzugang verzichten auf einen ausdrücklichen Fachbezug und enthalten stattdessen einen Bezug zu allgemeinen Kommunikationssituationen aus dem Studium. Dass Sprachtests für den Hochschulzugang ohne Fachbezug (im Sinne einer Fachsprache für ein bestimmtes Fachstudium) auskommen können, scheint Konsens zu sein. Dies ist auch die Vorgehensweise des TestDaF.

Sprachtests für den Hochschulzugang finden vor der Aufnahme des Studiums statt, daher wäre es – so lautet die in Kapitel 5.1 vorgestellte Argumentation der Kritiker – unangebracht, in einem Sprachtest für den Zugang zum Fachstudium Fachkenntnisse vorauszusetzen. Ebenso wie sich der Sprachenunterricht in der Studienvorbereitung vom studienbegleitenden Sprachenunterricht durch einen Verzicht auf einen Fachbezug unterscheidet, sollen sich demnach auch die Testverfahren vor der Aufnahme des Studiums eher an der allgemeinen Wissenschaftssprache orientieren. Lediglich in Sprachtests während des Studiums könnte eine Konzentration auf eine Fachsprache stattfinden. Eine andere Vorgehensweise wäre unfair. Diese Argumentation geht davon aus, dass Sprachtests mit Fachbezug nicht nur die Sprachkenntnisse, sondern auch Fachkenntnisse prüfen. Da Fachkenntnisse nicht als Teil des Testkonstrukts angesehen

werden, ist dies zu vermeiden. Fachkenntnisse können jedoch auch als notwendiger Bestandteil des Konstrukts von "Sprache im Studium" interpretiert und damit als konstitutives Element angesehen werden, welches den kompetenten Einsatz von Sprachkenntnissen erst ermöglicht und die Interpretation der Testergebnisse mit Blick auf die Sprachfähigkeit im Studium erleichtert.

Die Rolle der Fachkenntnisse ist ein zentrales Element der Diskussion um Sprachtests mit Fachbezug. In diesem Kapitel stelle ich daher Forschungsergebnisse zur Rolle der Fachkenntnisse vor und entwickle Fragestellungen für weiterführende Studien. In den meisten Studien wird die Diskussion um den Fachbezug in Sprachtests am Beispiel des Lesens geführt. Da auch in der vorliegenden Arbeit auf das Leseverstehen Bezug genommen wird, sollen Überlegungen zur Lesefertigkeit und zum Konstrukt von Leseverstehenstests vorgestellt werden.

Lesen und Leseverstehenstests: Einflussgrößen und Konstrukt

Studien zu Sprachtests mit Fachbezug konzentrieren sich häufig auf Tests zur Fertigkeit Lesen. Das ist plausibel, da Texte in universitären Lernsituationen ein wichtiges Kommunikationsmittel darstellen. Darüber hinaus bieten Tests zum Leseverstehen aus der Sicht der Testmethodik häufig den Vorteil einer mechanischen Auswertung. Die Aufgabentypen zum Leseverstehen sind in der Regel quantifizierbar, was eine objektive und reliable Bewertung von Leseverstehensaufgaben einfacher macht als beispielsweise eine Expertenbewertung von produktiven Tests zum Sprechen oder zum Schreiben.

Der Lesevorgang ist komplex. Modelle zum Lesevorgang leiden darunter, dass Lesen ein mentaler Prozess ist, der sich einer externen Beobachtung entzieht. Es gibt mehrere Modelle, welche den Leseprozess beschreiben. Man kann zunächst zwischen untergeordneten Fähigkeiten (Erkennen von Buchstaben und Wörtern) und übergeordneten Fähigkeiten unterscheiden. Das flüssige Lesen ist nach Grabe durch folgende Merkmale gekennzeichnet:

1. Reading is a rapid process.
2. Reading requires processing efficiency.
3. Reading requires strategic processing.
4. Reading is interactive.
5. Reading is purposeful.

6. Reading requires sufficient knowledge of language.
7. Reading requires sufficient knowledge of the world and of a given topic (Grabe, 1999: 12).

Verbreitet sind interaktive Modelle des Lesens, wobei sich "interaktiv" auf die Wechselwirkung zwischen Text und Leser bezieht. Stiefenhöfer erläutert:

> Lesen ist eine aktive Auseinandersetzung des Lesers mit dem vom Autor im Text versprachlichten Wissen. Im Verlauf der Textverarbeitung trägt der Leser sein in Form von Schemata organisiertes Sach- und Handlungswissen an den Text heran und verknüpft es mit den dort präsentierten Wissensstrukturen. [...] Lesen ist also kein bloßes Reagieren auf den Stimulus Text (Stiefenhöfer, 1995: 246-247).

Stiefenhöfer erwähnt die Schematheorie, welche die Diskussion um Modelle des Lesens beeinflusste. Nach der Schematheorie werden verfügbare Bestandteile des Wissens als Schema interpretiert. Auf diese Weise sollte eine Analyse des Verständnisprozesses ermöglicht werden. Aufgrund fehlender empirischer Belege und einer Vagheit des Begriffs regt Grabe an, die Schematheorie eher als Metapher denn als Abbild tatsächlicher Vorgänge beim Lesen anzusehen (zur Diskussion um Modelle des Lesens siehe Alderson, 1999; 2000; Alderson/Urquhart, 1984; Bernhardt, 1991b; 1999; Clapham, 1996; Grabe, 1991; 1999; 2002; Groeben, 1982; Grotjahn, 2000b; Lutjeharms, 1988; Samuels/Kamil, 1988; Urquhart/Weir, 1989).

Auch die Unterschiede zwischen dem Lesen in der Muttersprache und dem Lesen in der Fremdsprache sollen kurz skizziert werden: Beim Lesen in der Fremdsprache muss man mit einem geringeren Wortschatz auskommen, das Lesen in der Fremdsprache ist daher normalerweise langsamer und bezieht auch die formale Seite der Sprache mit ein. Häufig sind Leser in der Fremdsprache weniger mit den Eigenheiten authentischer Texte sowie mit dem Weltwissen vertraut, auf das im Text Bezug genommen wird (Alderson, 2000; Clapham, 1996; Grabe, 1999; Urquhart/Weir, 1989).

Der Unterschied zwischen der Lesefähigkeit in der Muttersprache und die Lesefähigkeit in der Fremdsprache ist in der Sprachwissenschaft häufig thematisiert worden. Besonderes Interesse gilt der Frage, unter welchen Bedingungen die Lesefähigkeit in der Muttersprache auf die Fremdsprache übertragen werden kann. Nach der von Clarke formulierten "short circuit hypothesis" kann auf die Lesefähigkeit in der Muttersprache nur zurückgreifen, wer in der Fremdsprache ein gewisses Niveau erreicht hat (Alderson, 1984; Bernhardt/Kamil, 1995; Clarke, 1980). Demnach ist die Lesefähigkeit in der Muttersprache behindert, wenn die Sprachkompetenz in der Fremdsprache zu niedrig ist. Die Lesefähigkeit kann nicht eingesetzt werden, sie wird "kurzgeschlossen". Die

Vorstellung, dass Leser in der Fremdsprache eine bestimmte sprachliche Schwelle in der Fremdsprachenkompetenz erreichen müssen, um ihre Lesefähigkeit der Muttersprache einsetzen zu können, wurde in einer Reihe von Studien erhärtet (Bernhardt/Kamil, 1995; Bossers, 1991; Carrell, 1991; Hacquebord, 1989, zit. in Bossers, 1991; Lee/Lemonnier-Schallert, 1997; Pichette/Segalowitz/Connors, 2003; Yamashita, 2001; 2004). Festzulegen, auf welchem Niveau der Fremdsprachenkenntnisse sich eine derartige Schwelle befindet, ist jedoch bislang nicht zufrieden stellend gelungen, da eine Vielzahl von Faktoren eine Rolle spielen: die Verständlichkeit des Textes oder das Niveau der geforderten Lesefertigkeit in der Muttersprache, um nur zwei zu nennen. Die "Kurzschlusshypothese" wird im Verlauf des Kapitels aufgegriffen, wenn ein vergleichbares Phänomen analysiert wird, die "Doppelte Schwellenhypothese" von Clapham (siehe Seite 220 ff).

Studien zum Einfluss der Lesefertigkeit in der Muttersprache auf die Lesefertigkeit in der Fremdsprache deuten außerdem darauf hin, dass Fremdsprachenkenntnisse einen wesentlich bedeutsameren Prädiktor für Leseleistungen in der Fremdsprache darstellen als Lesefertigkeit in der Muttersprache (Alderson, 1984; Bernhardt, 1991b; Bernhardt/Kamil, 1995; Bossers, 1991; Carrell, 1991; Hulstijn, 1991).

Ob die Lesefertigkeit auf einer Komponente beruht oder ob ihr tatsächlich mehrere Komponenten zugrunde liegen, wird in der sprachwissenschaftlichen Literatur kontrovers diskutiert (Rost, 1993; Weir/Huizhong/Yan, 2000). Die Argumentationen zur Dimensionalität des Leseverstehens verlaufen zum Teil ähnlich wie diejenigen zur Dimensionalität der Sprachkompetenz (siehe Kapitel 2.1, Seite 35 ff). Vertreter der These, dass sich die Lesefertigkeit auf eine einzige Komponente zurückführen lässt, argumentieren mit den Ergebnissen von Faktoranalysen. Häufig wurde beobachtet, dass Teilfertigkeiten stark auf einen Faktor laden, der als Lesefertigkeit interpretiert werden kann (Alderson, 1990; Rost, 1993; Urquhart/Weir, 1998). In anderen Studien ließ sich jedoch ein weiterer Faktor isolieren: Wortschatz (Weir/Porter, 1994). Wenn Lesefertigkeit wie von Stiefenhöfer oder Grabe als Produkt mehrerer Komponenten angesehen wird, beruhen die Argumentationen weniger auf statistischen Analysen als vielmehr auf einer Sammlung unterschiedlicher Variablen, die möglicherweise einen Einfluss auf die Lesefertigkeit haben. In Studien, die unter dieser Prämisse durchgeführt wurden, geht es darum, den Einfluss der Variablen unter unterschiedlichen Bedingun-

gen und mit unterschiedlichen Ausprägungen zu vergleichen. Wie bei der Diskussion um die Dimensionalität der Sprachkompetenz scheint sich die Ansicht der Mehrdimensionalität auch bei der Fertigkeit Lesen durchzusetzen (siehe Kapitel 2.1, Seite 35 ff). Auf der Seite des Lesers werden drei Hauptkomponenten herausgestellt: Sprachliche Fähigkeiten, strategische Fähigkeiten und Hintergrundwissen (Alderson, 2000; Grotjahn, 2000b; Urquhart/Weir, 1998).

Die Textrezeption setzt sprachliche und strategische Fertigkeiten voraus. Exemplarisch sei auf eine Liste Munbys verwiesen, die sprachliche Komponenten der Lesefertigkeit enthält:

- recognising the script of a language,
- deducing the meaning and use of unfamiliar lexical items,
- understanding explicitly stated information,
- understanding conceptual meaning,
- understanding the communicative value of sentences,
- understanding relations within the sentence,
- understanding relations between parts of text through lexical cohesion devices,
- understanding cohesion between parts of a text through grammatical cohesion devices,
- interpreting text by going outside it,
- recognising indicators in discourse,
- identifying the main point or important information in discourse,
- distinguishing the main idea from supporting details,
- extracting salient details to summarise (the text, an idea),
- extracting relevant points from a text selectively,
- using basic reference skills,
- skimming,
- scanning to locate specifically required information,
- transcoding information to diagrammatic display (Munby, 1987; zit. n. Alderson, 2000a: 10-11).

Es besteht jedoch weder Einigkeit darüber, welche Komponenten der sprachlichen Lesefertigkeit zuzuordnen sind, noch ist klar, ob sich einzelne Aspekte der sprachlichen Lesefertigkeit voneinander abgrenzen lassen (Alderson, 2000a: 32-84). Noch weniger Klarheit besteht mit Blick auf die strategischen Fähigkeiten:

> Thus, it is possible to say that reading strategies are important for comprehension, but also admit that it is an area of research which is not easy to categorize as a component process in any neat way, nor is it an area of reading research which has been well defined with respect to most of the issues discussed to this point (Grabe, 1999: 23).

Zur Lesefähigkeit zählt man bei der Annahme einer mehrdimensionalen Lesefertigkeit neben sprachlichen und strategischen Fertigkeiten auch Vorkenntnisse. Wie aus der Beschreibung des Lesevorgangs von Stiefenhöfer bereits hervorging, geht es beim Lesen nicht nur um die Rezeption eines Textes; der Leser bringt seine Vorkenntnisse vielmehr aktiv in den Leseprozess ein. Die neuen Informationen aus dem Text werden mit den vorhandenen Kenntnissen in eine Beziehung gebracht (Alderson, 2000a; Grabe,

1999; Stiefenhöfer, 1985; Urquhart/Weir, 1998). Bei fremdsprachlichen Texten hängt das Verständnis folglich nicht nur ab von den Sprachkenntnissen, sondern auch von den Vorkenntnissen zum Thema. Wie groß der Einfluss der Fachkenntnisse auf das Textverständnis ist, wird durch Modelle zum Lesevorgang nur unzureichend erläutert. Studien zur Rolle der Fachkenntnisse werden in diesem Kapitel vorgestellt (s. u.).

Vorkenntnisse können sich auf verschiedene Aspekte beziehen, die ich kurz erläutern werde. Bernhardt (1991b: 95-97) unterscheidet die Struktur des Wissens (*knowledge structure*) nach der Art und Weise, wie das Wissen erworben wird. Sie beschreibt drei Elemente: individuelles Wissen (*local-level knowledge*), Fachkenntnisse (*domain-specific knowledge*) und kulturelles Wissen (*cultural-specific knowledge*). Individuelles Wissen ergibt sich als Ergebnis aus individuellen Erfahrungen und Kenntnissen. Fachkenntnisse werden vor allem durch formelle Bildung erworben. Kulturelles Wissen wird als gemeinsamer kultureller Bestandteil von Gruppen von Generation zu Generation weitergereicht. Die Bedeutung dieser Unterscheidung erläutert Bernhardt an einer Studie mit 50 Katholiken aus Spanien und 50 Muslimen aus der arabischen Welt. Aus dem gemeinsamen kulturellen Hintergrund der Probandengruppen, das man durchaus erheben kann, ließen sich andererseits keine Rückschlüsse auf das individuelle Wissen bzw. auf die Fachkenntnisse ziehen (Bernhardt, 1991b: 97-117). Zu Vorkenntnissen zählt man in der Regel nicht Kenntnisse linguistischer Fachbegriffe. Studien zeigten, dass es keinen signifikanten Zusammenhang gibt zwischen der Vertrautheit mit sprachwissenschaftlichen Konzepten und der Sprachkompetenz in der Fremdsprache (Alderson/Clapham/Steel, 1997).

Weitere Variablen, welche einen Einfluss auf das Leseverstehen haben, sind situative Faktoren wie die Leseabsicht, die Stimmung sowie die Motivation und das Interesse der Leser (Alderson, 2000a; Urquhart/Weir, 1998; Weir/Huizhong/Yan, 2000).

Neben den Fähigkeiten und der Situation der Leser spielen auch Eigenschaften des Textes eine Rolle für das Leseverstehen. Texte können nach ihrer sprachlichen und inhaltlichen Komplexität unterschieden werden. Quantitative und qualitative Methoden zur Erfassung textueller Eigenschaften werden in Kapitel 6.1.2 (Seite 237 ff) vorgestellt und angewendet. In Zeiten des elektronischen Buchs sei schließlich darauf hingewiesen, dass auch die Präsentation des Texts den Verstehensprozess beeinflussen kann. Diesbezügliche Erkenntnisse sind für computerbasierte Tests von großem Interesse. Mit

Bezug auf den TOEFL wurde beispielsweise untersucht, inwieweit die Vertrautheit mit Computern einen Einfluss auf die Testleistungen hat (Taylor/Jamieson/Eignor/Kirsch, 1998; weitere Studien: Chalhoub-Deville, 1999; Grabe, 1999).

Tests zum Leseverstehen beziehen sich weniger auf den Lesevorgang, auf den Prozess des Lesens, sondern auf das Ergebnis dieses Prozesses: den Grad des Textverständnisses. Neben der Textauswahl kommt dabei den Aufgaben eine besondere Bedeutung zu. Das Textverständnis soll mithilfe von Fragen zum Text, Transferaufgaben oder anderen Aufgabentypen transparent werden. Der Aufgabentyp spielt eine zentrale Rolle dafür, welche Interpretationen das Testergebnis zulässt, welches Konstrukt dem Test also zugrunde liegt. Zu differenzieren ist zwischen unterschiedlichen Lesezielen und Lesestilen, die Gegenstand des Leseverstehenstests sind. Soll geprüft werden, ob die Kandidaten den Text in allen Einzelheiten verstehen? Dann wäre das detaillierte Lesen (auch: totales Lesen; engl.: *careful reading*) Gegenstand des Tests. Davon zu unterscheiden sind Leseziele wie "sich einen Eindruck verschaffen" (Lesestil: globales bzw. kursorisches Lesen; engl.: *skimming*), "eine gewisse Information finden " (Lesestil: suchendes bzw. selektives Lesen; engl.: *search reading* bzw. *scanning*) oder "Wichtiges und Unwichtiges im Text unterscheiden" (Lesestil: sortierendes bzw. orientierendes Lesen; engl.: *skimming*); (Lutjeharms, 1994; Weir, 1997; Weir/Huizhong/Yan, 2000; Westhoff, 1997). Weir, Huizhong und Yan (2000) kritisieren, dass Aufgaben in Leseverstehenstests zu häufig detailliertes Lesen verlangen und führen Beispiele aus IELTS und TOEFL an. Ausgehend von einer Analyse der Leseziele im Studium und der dafür benötigten Lesestile fordern sie, nicht nur das Detailverständnis zu prüfen, sondern auch das globale Verständnis und die Orientierung im Text zu berücksichtigen (auch Urquhart/Weir, 1998; Weir/Porter 1994).

Kann eine Ergebnisfeststellung mithilfe eines einzigen quantifizierbaren Faktors dem Leseverstehen angesichts der Komplexität des Testkonstrukts überhaupt gerecht werden?

> ... we will need to design and use a variety of reading assessment procedures to allow us to report on a variety of aspects of the student's ability to understand and to establish some systematic way of reporting the results on all of them. The differences this student shows across this range of results will inform us at least as much as will the result of adding them together. However good our tests are, a single score will always mislead (Spolsky, 1995: 151).

Diese von Spolsky vorgetragene Argumentation äußerte sich beispielsweise in der Weiterentwicklung von Deskriptoren, welche einzelne Ergebnisklassen genauer er-

läutern und dabei auf verschiedene Aspekte des Testkonstrukts eingehen. Zu nennen sind beispielsweise die Deskriptoren des Gemeinsamen Europäischen Referenzrahmens (Europarat/Rat für kulturelle Zusammenarbeit, 2001) oder der Ergebnisausweis beim TestDaF:

> TestDaF-Niveaustufe 3 (TDN 3):
> Kann sprachlich und inhaltlich einfach strukturierte schriftliche Texte zu hochschulbezogenen Themen in ihrem Gesamtzusammenhang und in ihren Einzelheiten verstehen und diesen Texten explizite Bedeutungen und Standpunkte entnehmen.
>
> TestDaF-Niveaustufe 4 (TDN 4):
> Kann sprachlich und inhaltlich komplex strukturierte schriftliche Texte zu hochschulbezogenen oder allgemein-wissenschaftlichen Themen in ihrem Gesamtzusammenhang und in ihren Einzelheiten verstehen und diesen Texten explizite Bedeutungen und Standpunkte entnehmen.
>
> TestDaF-Niveaustufe 5 (TDN 5):
> Kann sprachlich und inhaltlich komplex strukturierte schriftliche Texte zu allgemein-wissenschaftlichen Themen in ihrem Gesamtzusammenhang und in ihren Einzelheiten verstehen und diesen Texten sowohl explizite als auch implizite Bedeutungen und Standpunkte entnehmen (Projektgruppe TestDaF, 2000: 69-70).

Die Forderung Spolskys, verschiedene Testverfahren einzusetzen, wird ebenfalls von vielen Sprachtests umgesetzt, indem den Kandidaten unterschiedliche Texte zur Bearbeitung vorgelegt werden. Beim TestDaF müssen die Kandidaten beispielsweise nicht nur einen, sondern drei Texte mit unterschiedlichen sprachlichen Schwierigkeitsgraden und unterschiedlichen inhaltlichen Ausrichtungen lesen. Auf diese Weise steigt die Wahrscheinlichkeit, dass die komplexe Fähigkeit Leseverstehen umfassender abgebildet wird. Wenn die Variable "Vorkenntnisse" bewusst vernachlässigt werden soll, stellt der Einsatz unterschiedlicher Texte eine viel versprechende Möglichkeit der Testgestaltung dar. Ein offensichtlicher Nachteil ist die Verkürzung der Texte. Im Fachstudium werden häufig längere Texte (Bücher!) eingesetzt. Die Fähigkeit, sich in längeren Texten zurechtzufinden, lässt sich vermutlich besser an längeren Texten testen.

Ich möchte zusammenfassend auf folgende Zusammenhänge hinweisen: Lesen ist ein interaktiver Prozess, der nicht nur vom Verfasser des Texts abhängt, sondern auch von den Fähigkeiten, den Vorkenntnissen und der Interessenlage des Lesers. Zu den für das Verständnis wichtige Variablen gehören: die Person des Lesers (Sprachkompetenz, Strategien, Vorkenntnisse), die Situation (Leseziel, Leseinteresse) sowie Eigenschaften des Texts (inhaltlicher und sprachlicher Schwierigkeitsgrad, Struktur, Präsentation). Eine Abgrenzung der Variablen in der Praxis ist schwierig, der Einfluss einzelner Komponenten auf das Verständnis lässt sich daher nur ungenau ermitteln. Bei Tests zum Leseverstehen spielt nicht nur die Textauswahl, sondern auch der Aufgabentyp eine

zentrale Rolle, denn die Aufgaben legen fest, welche Lesestile eingesetzt werden müssen.

Studien zur Rolle der Vorkenntnisse

Zurück zum Lesen: Die Rolle der Vorkenntnisse in fachsprachlichen Tests wurde in einer Reihe von Studien untersucht, am häufigsten mit Blick auf das Leseverstehen (Alderson, 1988; Alderson/Urquhart, 1985a; 1985b; Alvermann/Hynd, 1989; Birjandi/Alavi/Salmani-Nodoushan, 2002; Chen/Graves, 1995; Erickson/Molloy, 1983; Hock, 1990; Hung, 1990; Koh, 1985; Osman, 1984; Peretz/Shoham, 1990; Ridgway, 1997; Shoham/Peretz/Vorhaus, 1987; Tan, 1990; zum Hörverstehen: Jensen/Hansen, 1995; zum Sprechen: Douglas/Selinker, 1993; Papajohn, 1999; zum Schreiben: Read, 1990; Tedick, 1990). Diese Studien kommen mehrheitlich zu dem Schluss, dass Vorkenntnisse einen signifikanten Einfluss auf das Leseverstehen haben. Allerdings sind die Forschungsergebnisse alles andere als eindeutig: Häufig erzielten Gruppen von Testteilnehmern nicht die erwarteten hohen Ergebnisse in Tests, obwohl sie über Vorkenntnisse verfügten. Ich stelle ausgewählte Studien vor.

Alderson und Urquhart (1985a; 1985b) untersuchten, ob ausländische Studenten in Leseverstehenstests mit fremdsprachlichen Texten aus ihrem eigenen Studienfach bessere Ergebnisse erzielen als mit Texten aus anderen Fachgebieten mit ähnlichem sprachlichen Anforderungsprofil. Die zugrunde liegende Hypothese war, dass Kandidaten mit Fachkenntnissen einen unbekannten Text besser erschließen können als Kandidaten ohne eine einschlägige Vorbildung. Alderson und Urquhart legten Teilnehmern aus studienvorbereitenden Sprachkursen, welche in der Regel bereits über ein abgeschlossenes Studium im Heimatland verfügten, fünf Texte aus verschiedenen Disziplinen vor. Sie mussten Fragen zum Text beantworten oder Lücken in diesem Text ersetzen. Sie kommen zu folgendem Ergebnis:

> The studies described in this paper have shown that academic background can have an effect on reading comprehension. They are thus a contribution to research into the nature of comprehension in general. They have also shown that particular groups of students may be disadvantaged by being tested on areas outside their academic field. If these findings are supported by further studies, then they will represent important evidence in support of the need for ESP [English for Specific Purposes] proficiency tests (Alderson/Urquhart, 1985b: 42).

Sie sehen mit der Studie ihre Hypothese im Grundsatz bestätigt, dass Fachkenntnisse in einem Test für den Hochschulzugang die Leistung in einem Leseverstehenstest mit einem Fachtext beeinflussen. Bei näherem Hinsehen erwiesen sich die Ergebnisse jedoch als nicht durchgehend eindeutig. Kandidaten mit wirtschaftswissenschaftlichem Hintergrund erzielten zwar bei wirtschaftswissenschaftlichen Texten signifikant bessere Ergebnisse als Kandidaten mit ingenieurwissenschaftlichem Hintergrund. Umgekehrt traf dieser Zusammenhang jedoch nicht zu, denn die Ökonomen erzielten beim Text zu Turbinen bessere Ergebnisse als die Ingenieure. Als die Studenten zusätzlich Leseverstehenstexte aus dem ELTS Test bearbeiten sollten, stimmten die Ergebnisse ebenfalls nicht vollständig mit der Hypothese überein: Die Ingenieur- und Naturwissenschaftler erzielten beim Techniktest bessere Ergebnisse als die Wirtschaftswissenschaftler, aber die Wirtschaftswissenschaftler waren beim Text aus den Sozialwissenschaften nicht wie erwartet besser als die Ingenieur- und Naturwissenschaftler. Es gab weitere Unstimmigkeiten, welche Alderson und Urquhart mit Blick auf den Schwierigkeitsgrad der Texte und den unterschiedlichen Sprachstand der Gruppen zu erklären versuchen. Sie räumen ein, dass ihre Erklärungsversuche die Phänomene nicht umfassend erläutern.

Die Studie von Alderson und Urquhart lässt Fragen offen:

- Die Zuschreibung der Fachkenntnisse scheint sehr allgemein gewesen zu sein. Die Gruppen wurden aufgrund des Studienfachs im Heimatland und dem Studienwunsch in Großbritannien eingeteilt. Es wurde nicht erhoben (oder nicht mitgeteilt), ob Kandidaten Fachkenntnisse über vermeintlich fachfremde Themen hatten. Auch ein Kandidat, der im Heimatland Psychologie oder Erziehungswissenschaften studierte, verfügt möglicherweise über Kenntnisse über Turbinen oder Elektrolyten – das waren Themen der "fachfremden" Texte. Vielleicht lernte er in der Schule darüber, vielleicht machte er praktische Arbeitserfahrungen damit oder machte sich zufällig zu einem dieser Themen kundig. Es ist nicht angemessen, nur bei Kandidaten aus ähnlichen Studienrichtungen bestimmte Fachkenntnisse vorauszusetzen. Eine individuelle Betrachtung der Lernbiografie der jeweiligen Kandidaten könnte zu einer differenzierteren Interpretation der einzelnen Ergebnisse führen. Alderson und Urquhart standen vor einem grundsätzlichen Problem, das auch in anderen Studien häufig nur unzureichend geklärt werden konnte: die Bestimmung der Art und Tiefe der Fach-

kenntnisse von Kandidaten in der Studienvorbereitung und die Bildung von "Fachgruppen".

- Ein weiterer Aspekt ist die Textschwierigkeit. Alderson und Urquhart bestimmen die Textschwierigkeit über die erzielten Ergebnisse. Haben alle Gruppen ein hohes Ergebnis erzielt, wird der Text als leicht verständlich eingeschätzt und umgekehrt. Diese Vorgehensweise berücksichtigt die Itemschwierigkeit nicht.

- Weiteren Erklärungsbedarf gibt es in Bezug auf den Sprachstand der einzelnen Kandidaten. Alderson und Urquhart nahmen eine Erhebung des Sprachstands durch Einstufungstests vor. Bei der Analyse der Ergebnisse gingen sie jedoch stets von den ursprünglich gebildeten Fachgruppen aus. Eine Gruppenbildung nach Sprachstand hätte möglicherweise weitere Erkenntnisse über den Zusammenhang zwischen dem Sprachstand und den Fachkenntnissen beim Leseverstehen von Fachtexten erbringen können. Da sich der Sprachstand der Fachgruppen im Mittel leicht unterschied, konnten einige Betrachtungen von Kandidaten mit unterschiedlichem Sprachstand angestellt werden. Doch die Ergebnisse sind nicht eindeutig, die einzig vertretbare Feststellung lautet: "Linguistic proficiency would clearly seem to be one factor involved" (Alderson/Urquhart, 1985b: 39).

In einer dritten Studie wurden Kandidaten Leseverstehenstests aus dem ELTS-Test mit und ohne Fachbezug zur Bearbeitung vorgelegt. Diese Studie lässt ebenfalls die Aussage zu, dass Fachkenntnisse einen Einfluss auf die Leistungen in Sprachtests mit Fachbezug haben, wobei die Ergebnisse ebenfalls einige Widersprüchlichkeiten aufweisen. Beim Technik-Modul des ELTS-Tests schienen die Fachkenntnisse einen besonders großen Einfluss auf die Verstehensleistungen zu haben. Alderson und Urquhart schlussfolgern, dass sich ein Fachbezug vor allem für Sprachtests für Ingenieure und Naturwissenschaftler eignet.

Neben der Aussage, dass Fachkenntnisse die Ergebnisse in Leseverstehenstests mit Fachtexten signifikant beeinflussen – wenn auch nicht durchgehend –, haben die Studien von Alderson und Urquhart wichtige methodische Fragen aufgeworfen. So muss man offensichtlich auch den Sprachstand der Kandidaten und den Fachlichkeitsgrad der Texte genauer beschreiben, wenn man Aussagen zum Verhältnis von Fach-

kenntnissen und Sprachkenntnissen treffen will. Ein lineares Verhältnis scheint nicht zu bestehen (Alderson/Urquhart, 1985a; 1985b).

Aus ihren Untersuchungen, mit denen die Einflüsse von Fachkompetenz und Sprachkompetenz in Leseverstehenstests mit Fachtexten erfasst werden sollten, zogen Alderson und Urquhart (1985a; 1985b) darüber hinaus Schlüsse über die Einteilung der Kandidaten in Fachgruppen. Wie viele Texte benötigt man in einem Sprachtest für den Hochschulzugang? Benötigen angehende Studenten der Elektrotechnik andere Texte als Maschinenbauer? Wie soll man mit Studenten aus kombinierten Studiengängen, die immer populärer werden, umgehen? In der ersten Studie wurden die Teilnehmer in vier Gruppen eingeteilt: In die erste Gruppe gehörten Kursteilnehmer, welche in der Mehrzahl bereits ein wirtschaftswissenschaftliches Studium abgeschlossen hatten und die ein Aufbaustudium anstrebten ("W-Gruppe"). Die zweite Gruppe bestand aus Teilnehmern, die bereits ein ingenieurwissenschaftliches Studium absolviert hatten und die nun unterschiedliche technische Aufbaustudiengänge belegen wollten ("T-Gruppe"). In der dritten Gruppe befanden sich Studenten der Mathematik oder der Physik ("MP-Gruppe"). Die vierte Gruppe umfasste Studenten geisteswissenschaftlicher Studiengänge (Erziehungs-, Sprachwissenschaften, Psychologie usw., "PS-Gruppe"). Allen Gruppen wurden fünf Fachtexte aus unterschiedlichen Fachgebieten zur Bearbeitung vorgelegt. Die Ergebnisse der PS-Gruppe waren dabei am höchsten. Es stellte sich heraus, dass die Ergebnisse der T-Gruppe und die Ergebnisse der MP-Gruppe in den unterschiedlichen Texten jeweils ähnlich waren. Vergleichbare Ergebnisse erzielten auch die Kandidaten aus der W-Gruppe und aus der PS-Gruppe. Alderson und Urquhart schlussfolgern:

> It is interesting that, with minor differences, engineering and mathematics/physics students can perhaps be regarded as forming two closely related groups. Similarly, there are close resemblances between the Administration/Finance group and the Liberal Arts group (Alderson/Urquhart, 1985b: 32).

Diese Ergebnisse konnten in einer zweiten Studie bestätigt werden. Es wurden ebenfalls vier Gruppen gebildet, die den Gruppen aus der ersten Studie ähnelten. Diese Beobachtungen legen den Schluss nahe, dass es nicht unbedingt notwendig ist, für jedes Studienfach einen gesonderten Text auszuwählen, wenn man einen Leseverstehenstest mit Fachbezug in einem Test für den Hochschulzugang konzipieren möchte. Für die Kandidaten hätten zwei Texte ausgereicht. Diese Aussage geschieht unter der Einschränkung, dass relativ wenig über den Grad der Fachlichkeit und das Schwierigkeitsniveau der

verwendeten Texte mitgeteilt wird. In dem Bericht über die Studie wird lediglich mitgeteilt, dass die Texte in Bezug auf die Satzlänge und die Länge der Wörter vergleichbar waren. Es ist denkbar, dass eine Reduktion auf (in diesem Fall) zwei Texte einige Kandidaten benachteiligt, wenn der Fachlichkeitsgrad der Texte stärker ausgeprägt ist.

Tan (1990) konnte die Leistungen im Verständnis von fachspezifischen Texten durch die Fachkenntnisse und durch die Sprachkenntnisse vorhersagen. Sie untersuchte die Leistungen im Leseverstehen von malaysischen Studierenden im dritten Studienjahr aus den Fachbereichen Medizin, Jura und Wirtschaftswissenschaften und stellte fest, dass sich die Kandidaten beim Lesen fremdsprachiger Texte sowohl ihrer Fachkenntnisse als auch ihrer Fremdsprachenkenntnisse bedienen. Wenn Texte aus dem eigenen Studienfach gelesen wurden, waren Sprachkenntnisse und Fachkompetenz signifikante Prädiktorvariablen, wobei sich Sprachkenntnisse als wesentlich bedeutsamere Prädiktorvariable erwiesen. Wenn es um fachfremde Texte ging, nahm die Bedeutung der Sprachkompetenz für das Verständnis deutlich zu. In Bezug auf die Fachkompetenz schlussfolgert Tan, dass sie über einer bestimmten Schwelle liegen muss, um beim Verständnis von Fachtexten zu helfen. Eine Differenzierung, bei welchem Niveau der Sprachkenntnisse dieser Effekt besonders stark auftritt, wird nicht vorgenommen (Tan, 1990, 214-224).

In der Studie von **Hammadou (1991; 2000)** ging es um die Frage, ob Fachtexte, in denen das Thema durch Analogien verdeutlicht werden soll, für Leser verständlicher sind. Dabei wurden Leser in der Muttersprache und in der Fremdsprache berücksichtigt. Da die Probanden der Studie (französische und amerikanische Studenten) auch nach ihrem Sprachstand und ihren Vorkenntnissen differenziert wurden, gewann Hammadou zusätzlich Erkenntnisse über den Einfluss der Fachkenntnisse auf den Verstehensprozess. Auch Hammadou stellte eine Korrelation zwischen Fachkenntnissen und Textverstehen fest. Doch die Sprachkenntnisse zeigten sich in seiner Studie ebenfalls als weitaus wichtigerer Faktor. Auch Kandidaten, denen das Thema des Textes unbekannt war, erzielten bei der Aufgabe, die Informationen aus einem Fachtext aus der Erinnerung aufzuschreiben, gute Ergebnisse (Hammadou, 1991; 2000).

Auf vergleichbare Ergebnisse treffen **Jensen und Hansen (1995)** mit Bezug auf das Hörverstehen. Sie untersuchten die Verstehensleistungen von unterschiedlichen Vorlesungen. Die Kandidaten hörten einen "nichtakademischen" Vortrag und mehrere

Fachvorträge. Ihre Verstehensleistungen wurde anhand von Fragen zum Inhalt untersucht. Methodisch unterscheidet sich ihr Vorgehen von demjenigen in Aldersons und Urquhardts Studie: Sie erhoben systematisch, ob die Kandidaten Vorkenntnisse zu dem Thema besaßen oder nicht, und bestimmten diese nicht einfach aufgrund des belegten Studienfachs. Jensen und Hansen stellten mit Hilfe von multiplen Regressionsanalysen fest, dass die Leistungen im allgemeinen Vortrag ein besserer Prädiktor auf die Leistungen im Verstehen der Fachvorträge waren als Fachkenntnisse. Wenn ein Effekt der Fachkenntnisse auf die Verstehensleistungen zu beobachten war, dann handelte es sich um Vorlesungen aus technischen Disziplinen. Allerdings war der Effekt der Fachkenntnisse in jedem Fall geringer als derjenige der Sprachkenntnisse. Zwei Informationen fehlen, um die Aussagekraft der Studie bestimmen zu können: Aus dem Bericht über die Studie geht nicht hervor, wie ausgeprägt der Fachlichkeitsgrad der Vorlesungen war. Außerdem wird wenig über das Niveau der Sprachkenntnisse mitgeteilt.

Eine Studie zum Sprechen kommt zu dem Ergebnis, dass auch bei der Sprachproduktion sowohl Fachkenntnisse als auch Sprachkenntnisse eingesetzt werden, die Sprachkenntnisse jedoch eine größere Rolle spielen. **Douglas und Selinker (1993)** vergleichen die Leistungen von zwölf fremdsprachigen Studenten mit einem ersten Studienabschluss in Mathematik aus ihrem Heimatland in zwei Tests zum Sprechen. Mit den Tests sollte die Eignung der Kandidaten für die Tätigkeit als Lehrassistenten festgestellt werden. Zunächst unterzogen sich die Kandidaten dem "*Speaking Proficiency English Assessment Kit*" (SPEAK), einem Sprachtest zur mündlichen Sprachproduktion ohne Fachbezug, der aus dem "*Test of Spoken English*" (TSE) hervorgegangen ist. Analog zum SPEAK entwickelten Douglas und Selinker in Zusammenarbeit mit einem Mathematiker eine fachspezifische Version mit Bezug zur Mathematik, den sie als MATHSPEAK bezeichneten. Diesem Test unterzogen sich die Kandidaten mit einem zeitlichen Abstand. Im Gegensatz zu Tests zum Lesen oder zum Hören, die in der Regel eine objektive Bewertung ermöglichen, beruhen die Ergebnisse beim SPEAK bzw. MATHSPEAK auf subjektiven Bewertungen, so dass das Problem der Reliabilität zwischen unterschiedlichen Prüfern in ihrer Untersuchung eine besondere Bedeutung erhält. Douglas und Selinker berichten von einigen Schwierigkeiten mit der Reliabilität zwischen unterschiedlichen Prüfern (*interrater reliability*) und auch zwischen den beiden Testversionen. Mit Blick auf die Ergebnisse stellten sie fest, dass sich die Ergebnisse der Kandidaten in beiden Tests stark unterschieden. Wenn sie den

Schwellenwert bei beiden Tests gleich ansetzen, kommt es zu folgenden Unterschieden: Vier Studenten, welche beim SPEAK durchgefallen waren, hätten MATHSPEAK bestanden. Drei Studenten, welche den SPEAK bestanden hatten, wären beim MATHSPEAK durchgefallen. Die übrigen fünf Kandidaten wären entweder bei beiden Tests durchgefallen oder hätten beide bestanden. Eindeutig waren lediglich einige Teilergebnisse: In den Kategorien Grammatik und "fluency" erzielten die Kandidaten im MATHSPEAK durchweg bessere Ergebnisse als im SPEAK. Douglas und Selinker schlussfolgern:

> [T]he grammar finding in particular suggests to us that there is more to taking a specific purpose test than simply background knowledge. At least as perceived by raters, the differences in contextual method (as embodied in method facets) between the SPEAK and MATHSPEAK are related to differences in the way language competence is realized (Douglas/Selinker, 1993: 243-244).

Die Studie von Douglas und Selinker gibt insgesamt keine schlüssige Antwort auf die Frage, ob in diesem Fall Tests mit Fachbezug oder Tests ohne Fachbezug vorzuziehen wären. Sie unterstützt jedoch die Einschätzung, dass Fachkenntnisse als Teil der Sprachkompetenz komplexe sprachliche Äußerungen ermöglichen.

Read (1990) und **Tedick (1990)** führten Studien zum Schreiben durch. Tedick ließ fremdsprachige Studierende zwei unterschiedliche Schreibaufgaben bearbeiten. Ein Thema stammte aus ihrem Studienfach, das andere Thema war allgemeiner Natur. Die Texte zu dem Thema aus dem Studienfach waren von höherer Qualität. Sie stellte außerdem fest, dass die fachlichen Texte stärker zwischen den Kandidaten differenzierten, und schlussfolgert, dass die Verwendung von Themen mit Fachbezug ein geeigneteres Instrument zur Bewertung des Schreibens im Studium darstellt als allgemeine Themen. Read kommt in einer Analyse von drei Schreibaufgaben jedoch zu dem Schluss, dass unterschiedliche Aufgabenstellungen die Rangfolge der Kandidaten nicht signifikant verändern.

Die bislang vorgestellten Studien zur Rolle von Fachkenntnissen für den Leseverstehensprozess ergeben kein einheitliches Bild. Gemeinsame Ergebnisse sind: Für das Verständnis von Fachtexten spielen Sprachkenntnisse eine wichtige Rolle. Eine Besonderheit scheint der Umgang mit Fachtexten aus technischen oder naturwissenschaftlichen Fachgebieten zu sein. Kandidaten mit naturwissenschaftlicher oder technischer Vorbildung erzielen häufig bessere Ergebnisse. Alle Studien sind auf zwei Schwierigkeiten gestoßen: Zum einen war es schwierig, das Niveau der Fachkenntnisse

einzelner Kandidaten zu bestimmen, was Aussagen über die Rolle der Fachkenntnisse beim Leseverstehensprozess erschwert. Zum anderen ist davon auszugehen, dass sich der Grad der Fachbezogenheit der einzelnen Texte stark unterscheidet bzw. nicht genau beschrieben wurde. Daher sind weiter gehende Aussagen über das Verhältnis von Fachkenntnissen und Sprachkenntnissen beim Leseverstehen von Fachtexten auf der Basis dieser Studien nicht möglich.

Ich stelle noch zwei weitere Studien vor: **Fox, Graves, Jennings und Shohamy (1999)** gehen bei der Untersuchung des Canadian Academic English Language (CAEL) Tests, einem Test für den Nachweis von Englischkenntnissen für den Hochschulzugang an englischsprachigen Universitäten in Kanada, nicht vom Fachbezug aus, sondern von der Frage, ob eine Auswahl aus verschiedenen Themen dazu führt, dass die Kandidaten bessere Ergebnisse erzielen. Untersucht wurden die Ergebnisse aus den Prüfungsteilen *Reading* (Leseverstehen), *Lecture* (Hörverstehen), *Essay* (Textproduktion) sowie die Gesamtergebnisse. Die erste Gruppe konnte aus den folgenden Themen eines auswählen: *Marine Animals*, *Food*, *Forests* und *Weather*. Die Themen wurden in allen Prüfungsteilen aufgegriffen. Eine zweite Gruppe musste einen Test zu einem vorgegebenen Thema bearbeiten. Die erste Gruppe erzielte zwar durchschnittlich etwas bessere Ergebnisse als die Gruppe, welche nicht auswählen konnte, die Unterschiede waren jedoch nicht signifikant. In diesem Fall hatte das Thema und damit auch die fachlichen Vorkenntnisse keine signifikante Auswirkung auf das Testergebnis (Fox/Graves/Jennings/Shohamy, 1999). Die Studie unterstreicht den großen Einfluss der Sprachkenntnisse auf die Ergebnisse in Sprachtests mit Fachbezug. Ob sich ein signifikanter Einfluss der Vorkenntnisse bei mittlerer Sprachkompetenz feststellen ließ, wurde nicht ausdrücklich untersucht; es kann also auch nicht ausgeschlossen werden. Die Studie kommt zu dem Schluss, dass Sprachtests mit erkennbarem Fachbezug die Testfairness nicht beeinträchtigen müssen.

Die Studie von **Papajohn (1999)** bietet auch Erfahrungen zur Ausgestaltung von Sprachtests mit Fachbezug. Nicht immer können Tests, welche gemeinsprachliche Fertigkeiten prüfen, und Tests, welche fachsprachliche Fertigkeiten prüfen, eindeutig unterschieden werden.

> It is important to note that tests are not *either* general purpose *or* specific purpose; there is rather a continuum of specificity from very general to very specific, with a given test falling at any point on the continuum (Douglas, 1997: 111).

Die Wahl des Themas und der Texte ist ein wichtiger Faktor, der Auswirkungen auf die Konstruktvalidität des Tests hat. Dies war ein zentrales Ergebnis der Studie von Clapham. Ähnliche Beobachtungen machte Papajohn am Beispiel des TEACH-Tests ("Taped Evaluation of Assistants' Classroom Handling"). Papajohn untersuchte die Auswirkungen unterschiedlicher Themen innerhalb eines Fachgebiets am Beispiel eines Tests zur mündlichen Sprachproduktion mit Fachbezug. Seine Ergebnisse sind im Zusammenhang dieser Arbeit von Interesse, obwohl es in dem Test um die gesprochene Sprache geht. Sie verdeutlichen die Schwierigkeiten, selbst innerhalb eines Fachgebiets mehrere Tests mit gleichem Schwierigkeitsgrad zu entwerfen.

Der an der *University of Iowa* entwickelte TEACH Test ("*Taped Evaluation of Assistants' Classroom Handling*") ist ein fachsprachlicher Test zum mündlichen Ausdruck, der sich an ausländische Lehrassistenten richtet (Abraham/Plakans, 1990; Papajohn, 1999). Von diesem Test gibt es mehrere Versionen zu unterschiedlichen Fachgebieten. Er kann als Test mit Fachbezug par excellence gelten: ein Performanztest, bei dem Inhalt und Items aus einer Analyse der Sprachverwendungssituation im Fach gewonnen wurden. Die Grenzen zwischen einem beruflichen Eignungstest und einem Sprachtest lassen sich nicht mehr eindeutig ziehen.

Beim TEACH müssen die Kandidaten eine kurze Lehreinheit aus ihrem Fachgebiet durchführen. Sie erhalten einen Tag vor ihrer Prüfung ein Kapitel aus einem Fachbuch zum jeweiligen Fachgebiet. Die Themen und das Material zur Vorbereitung werden von Lehrkräften aus dem jeweiligen Fachbereich ausgewählt. Die Prüfung besteht aus zwei Teilen: einem kurzen Vortrag (fünf Minuten) zu einem Thema, auf das sie sich einen Tag lang vorbereiten können und aus Fragen der Zuhörer die beantwortet werden müssen. Die Zuhörer setzten sich aus zwei Prüfern und einigen Studenten zusammen, die Prüfer stellen das Ergebnis auf einer Skala von null bis vier fest ("*not competent – not adequate – minimally adequate – competent*"). Das Ergebnis aus dem TEACH entscheidet jedoch nicht allein über den Einsatz des Kandidaten als Lehrassistent. An der *University of Iowa* müssen sich fremdsprachige Lehrassistenten zusätzlich dem SPEAK Test ("*Speakting Proficiency English Assessment Kit*") unterziehen, einem allgemeinsprachlichen Test, der aus einer 20-minütigen mündlichen Prüfung besteht. Dieser Test wurde vom "*Educational Testing Service*" konzipiert und inzwischen vom "*Test of Spoken English*" (TSE) abgelöst. Am Rande sei darauf hingewiesen, dass man einen

Sprachtest ohne Fachbezug mit einem Sprachtest mit Fachbezug kombiniert. Man ist offensichtlich der Ansicht, dass die Ergebnisse aus einem Sprachtest allein für die Entscheidung über den Einsatz nicht ausreichen.

Innerhalb des TEACH-Tests mit Bezug zur Chemie untersuchte Papajohn den Einfluss der unterschiedlichen Themen auf die Leistung der Kandidaten. Als Kontrollinstrument setzte er den SPEAK-Test ein. Bei Themen, die er als inhaltlich einfach bzw. komplex einstufte, ergaben sich signifikante Unterschiede in der bewerteten Leistung der Kandidaten. Er resümiert:

> This suggests that using field-specific tests for examinees does not automatically take care of making topics equivalent. Test developers should analyse topic features and compare topics, not only regarding the theme of the topic, like chemistry, but also comparing topic features, like concepts, maths and calculations (Papajohn, 1999: 75).

Diese Untersuchung stellt nicht das Projekt Sprachtests mit Fachbezug in Frage. Sie dämpft aber die Erwartung, dass Sprachtests mit Fachbezug zumindest innerhalb eines Fachgebietes über eine hohe Paralleltestreliabilität verfügen oder eine einfachere Themenauswahl bieten. Umso schwieriger ist es, mehrere Tests mit Bezügen zu unterschiedlichen Fachgebieten zu entwickeln, deren Anforderungsniveaus vergleichbar sind. Dies verdeutlicht die Notwendigkeit, Tests zu erproben. Dies ist beim Leseverstehen eher möglich als beim Sprechen, im Falle der DSH jedoch eher die Ausnahme.

Clapham (1996; 2000) verschaffte sich im Rahmen einer breit angelegten Studie nicht nur einen genauen Eindruck vom Sprachstand der Kandidaten, sondern berücksichtigte auch die Lesegewohnheiten und Fachkenntnisse. Ein wichtiges Element ihrer Studie war auch die Beachtung der Textschwierigkeit. Clapham führte die Studie am Beispiel des Prüfungsteils Leseverständnis von IELTS durch. Bis zur Überarbeitung im Jahre 1995 enthielt der IELTS-Test Lesetexte mit Fachbezug. Die Textes stammten aus folgenden Disziplinen: Biologie, Physik, Sozialwissenschaften und Geisteswissenschaften. Mit ihrer Studie hat Clapham den Kenntnisstand über Sprachtests mit Fachbezug systematisiert und erweitert.

Eine erste Beobachtung bezog sich auf das Verhältnis des Studienfachs auf die Leistungen in Leseverstehenstests mit Fachbezug. Bei ihrer Untersuchung stieß sie auf die Schwierigkeit, die Studierenden bestimmten Fachgebieten zuzuordnen. Anders als bei ausgebildeten Fachkräften mit Berufserfahrung sind die Fachkenntnisse und Fachsprachenkenntnisse bei Studienbewerbern noch nicht so stark ausgeprägt. Häufig hatten

sich die Studienbewerber mehr oder weniger zufällig bereits mit dem Thema eines Textes beschäftigt, das nicht aus dem angestrebten Studienfach stammte. Demnach waren auch die Ergebnisse uneinheitlich. Zwar erzielten in der Hauptstudie die Kandidaten insgesamt signifikant bessere Ergebnisse, wenn die Texte aus dem eigenen Studienfach stammten. In der Pilotstudie wurden allerdings keine signifikant besseren Ergebnisse erzielt. Möglicherweise ist es zu diesem Unterschied gekommen, weil der Fachlichkeitsgrad der Texte unterschiedlich stark ausgeprägt war. Sie untersuchte auch den Unterschied zwischen den Ergebnissen beim Leseverstehen von Studierenden vor dem ersten akademischen Grad (*undergraduate students*) und von Studierenden im Aufbaustudium (*postgraduate students*). Es stellte sich heraus, dass das Ergebnis im Leseverstehen bei den "*undergraduates*" relativ unabhängig vom Fachlichkeitsgrad der Texte war. Bei den "*postgraduates*" stellte sie jedoch einen Einfluss des Fachbezugs fest.

Eine zweite Beobachtung bezog sich auf den Fachlichkeitsgrad der Texte. Clapham hatte keinen Einfluss auf die Auswahl der Texte genommen, sondern sich an diejenigen gehalten, welche von den IELTS-Testentwicklern in Zusammenarbeit mit Experten aus den Fächern ausgewählt worden waren. Sie stellte fest, dass sich der Fachlichkeitsgrad der Texte erheblich unterschied. Selbst bei einem standardisierten Test, der in einem aufwändigen Verfahren unter Beteiligung von Testmethodikern, Linguisten und Fachkräften der betroffenen Fächer erstellt wird, kann offensichtlich ein mehr oder weniger einheitliches Niveau des Fachlichkeitsgrads kaum erreicht werden. Dies überrascht zunächst, denn zur Ermittlung der Spezifik von Fachsprachen sind von der Fachsprachenforschung konkrete Methoden formuliert worden. Hoffmann (1985) erläutert beispielsweise, nach welchen Kriterien sich eine vertikale Schichtung der Fachsprachen vornehmen lässt und wie sich die Merkmale von Fachtexten in einer kumulativen Analyse erheben lassen. Er ist der Überzeugung, dass man die Spezifik der Fachsprachen "recht gut und relativ einfach mit quantitativen Methoden erfassen" kann (1985: 243). Dennoch führen diese Methoden nicht zu einer zufrieden stellenden Abbildung des Merkmals Textschwierigkeit (im Zusammenwirken mit der Aufgabenschwierigkeit), wie sie für Sprachtests nötig wäre. Ursache dürfte die Komplexität der Merkmale Textschwierigkeit und Aufgabenschwierigkeit sein. Im Nachhinein ist der Grad der Fachlichkeit und die Textschwierigkeit zu erheben, wenn man ein bereits über ein Bild von den Fachkenntnissen und den Sprachkenntnissen der Kandidaten verfügt.

Clapham schlussfolgerte aus ihrer Studie, dass der Einfluss der Fachkenntnisse für das Leseverstehen mit steigendem Grad der Fachlichkeit zunimmt. Sie machte diesen Unterschied an den Eigenschaften der Texte fest und an den Ergebnissen der Kandidaten. Mit Blick auf die Texteigenschaften stellt sie heraus, dass der Fachlichkeitsgrad von Passagen, die in ein Thema einführen oder welche Forschungsergebnisse präsentieren, geringer ist als derjenige von Passagen, die bestimmte Abläufe beschreiben und analysieren.

Schließlich untersuchte Clapham die Intensität, mit der Fachkenntnisse und Sprachvermögen auf die Ergebnisse beim Verständnis von Texten unterschiedlicher fachsprachlicher Herkunft wirken. Dazu analysierte sie die Ergebnisse aus Leseverstehenstests mit mehr oder weniger ausgeprägtem Fachbezug. Während Sprachkenntnisse allein bereits 44 Prozent der Variation der Ergebnisse erklärten, erhöhte die Variable Fachkenntnisse diesen Wert nur um einen Prozent. Diese Beobachtung legt nahe, dass die Leistungen im Leseverstehen in sehr viel größerem Maße von den Sprachkenntnissen abhängen und nur zu einem geringen Anteil von den Fachkenntnissen, dass die Sprachkenntnisse also ein besserer Prädiktor für Leistungen im Leseverstehen von Fachtexten darstellen als Fachkenntnisse. Diese Beobachtung stimmt mit denen der oben erwähnten Studien überein. In einem zweiten Schritt wurden Texte, deren fachsprachlicher Charakter nur schwach ausgeprägt war, von der Analyse ausgeschlossen. Als nur Texte mit einem hohen Fachlichkeitsgrad in die Analyse einbezogen wurden, war der Beitrag der (durch den Grammatiktest gemessenen) Sprachkompetenz mit 26 Prozent immer noch deutlich. Überraschend war, dass eine Berücksichtigung der Fachkompetenz als Prädiktorvariable diesen Wert auf 38 Prozent steigerte. Daraus folgert Clapham, dass die Fachkenntnisse bei zunehmendem Fachlichkeitsgrad der Texte eine zunehmend wichtige Rolle für das Verständnis spielen.

Mit dem Fachlichkeitsgrad und dem sprachlichen Schwierigkeitsgrad untersucht Clapham zwei Aspekte, welche in anderen Untersuchungen häufig nur wenig Beachtung fanden. Laut Clapham ist nicht allein die Anzahl des Fachwortschatzes Ausschlag gebend für die Schwierigkeit eines Fachtextes, sondern die jeweilige Einführung der Fachwörter. Mit Bezug auf die Textsorte "Forschungsberichte" machte Clapham bei verschiedenen Abschnitten der Texte unterschiedliche Schwierigkeitsgrade aus. In der Einleitung von Forschungsberichten wird demnach weniger Fachsprache verwendet als

in der Beschreibung von Forschungsgegenständen. Weniger als die Quelle spielt also die kommunikative Funktion des jeweiligen Abschnittes eine Rolle für den (fach-)sprachlichen Schwierigkeitsgrad.

Was ist aus den Studien für die Gestaltung von Sprachtests für den Hochschulzugang zu folgern? Die folgenden Aspekte möchte ich hervorheben: Erstens die Beobachtung von Papajohn (1999) zur Äquivalenz von Sprachtests mit Fachbezug; zweitens die Feststellung von Alderson und Urquhart (1985a; 1985b) zur Anzahl der notwendigen Fachgebiete, die von Sprachtests mit Fachbezug abgedeckt werden müssen; drittens methodische Fragen zur Erhebung der Sprachkenntnisse und viertens die für diese Studie wichtigen Informationen zum Verhältnis von Vorkenntnissen und Sprachkenntnissen bei Sprachtests mit Fachbezug.

Von Belang ist meiner Ansicht nach die Beobachtung von Papajohn (1999), der – freilich mit Blick auf das Sprechen – feststellte, dass ein Fachbezug nicht unweigerlich zu einer hohen Äquivalenz von Sprachtests führt. Bei Sprachtests mit Fachbezug ist eine hohe Vergleichbarkeit demnach nicht eher gegeben als bei Sprachtests, die einen Fachbezug vermeiden.

Zur Ökonomie von Sprachtests mit Fachbezug sind die Beobachtungen von Alderson und Urquhart von Interesse: Sie stellten fest, dass ein Fachbezug nicht für jedes Studienfach hergestellt werden muss. Die Anzahl der Texte ist zwar höher als bei Sprachtests ohne Fachbezug, sie muss aber nicht ins Unbewältigbare steigen: Selbst bei vier unterschiedlichen Versionen ergaben sich bereits Überschneidungen, obwohl die Studenten, die an der Studie teilnahmen, aus ganz unterschiedlichen Disziplinen kamen.

An den Studien von Alderson und Urquhart wurde darüber hinaus ein methodisches Problem deutlich: Wie können die Vorkenntnisse der Kandidaten erfasst werden? Durch eine systematischere Erhebung der Vorkenntnisse konnten Jensen und Hansen (1995) oder Clapham (1996) den Kenntnisstand erweitern. Verschiedene Verfahren sind zur Erhebung der Vorkenntnisse nötig.

Auch wenn die Ergebnisse der Untersuchungen zur Rolle der Vorkenntnisse nicht völlig kongruent sind, so lässt sich feststellen: Das Niveau der Sprachkenntnisse hat beim Lesen fremdsprachiger Texte einen deutlich größeren Einfluss auf das Verständnis als eventuell vorhandene Vorkenntnisse.

Doppelte Schwellenhypothese: Interaktion zwischen Vorkenntnissen und Sprachkenntnissen

Vergleichbar mit der "short circuit hypothesis" von Clarke, die sich auf die Lesefertigkeit in der Muttersprache und ihre Übertragung in die Fremdsprache bezieht, formulierte Clapham die "Doppelte Schwellenhypothese". Diese Hypothese beschreibt, wie sich der Einfluss der Vorkenntnisse beim Lesen in der Fremdsprache in Abhängigkeit von dem Niveau der Fremdsprachenkenntnisse verändert. Clapham beobachtete folgende Entwicklung: Es existiert eine sprachliche Schwelle, die erreicht werden muss, damit Leser in der Fremdsprache ihre Vorkenntnisse zum Thema überhaupt nutzen können. Clapham beobachtete eine weitere Schwelle, bei der sich der Einfluss der Vorkenntnisse wieder änderte: Leser mit besonders fortgeschrittenen Fremdsprachenkenntnissen können fremde Texte offensichtlich auch unabhängig von Vorkenntnissen erschließen. Vor allem bei Lesern mit "mittleren" Fremdsprachenkenntnissen ist – so lautet das Fazit von Clapham – der Einfluss der Vorkenntnisse besonders ausgeprägt.

Clapham war nicht die erste, die auf einen Zusammenhang zwischen Vorkenntnissen, Fremdsprachkenntnissen und Leistungen im Leseverstehen stieß. Laufer und Sim stellten beispielsweise mit Blick auf einen notwendigen Wortschatzumfang fest, dass Schüler ihre Lesefertigkeit in der Muttersprache nicht unabhängig von dem Niveau der Fremdsprachenkenntnisse in einem fremdsprachlichen Text einsetzen konnten. Sie schlussfolgern, dass das Leseverstehen von bestimmten fremdsprachlichen Texten erst dann möglich ist, wenn ein bestimmter Wortschatz vorhanden ist. Liegt der Wortschatz unter einer bestimmten Schwelle, können Vorkenntnisse oder Lesefertigkeit in der Muttersprache nicht eingesetzt werden (Laufer/Sim, 1985). Hudson machte die Beobachtung, dass Fremdsprachenlerner auf eine inhaltliche Vorbereitung auf den Text in Abhängigkeit von ihren Vorkenntnissen unterschiedlich profitieren (Hudson, 1982). Alderson beschreibt eine "Schwellenhypothese" (*threshold hypothesis*), die sich jedoch nur auf die untere Schwelle bezieht (Alderson, 1984). Das Besondere an Claphams Doppelter Schwellenhypothese ist die (scheinbar) breite empirische Basis sowie die Feststellung, dass es eine weitere Schwelle gibt, ab der ein Textverständnis relativ unabhängig von Vorkenntnissen zustande kommt.

Im Folgenden werde ich die Forschungsergebnisse von Clapham (1996; 2000) und Ridgway (1997) zur Doppelten Schwellenhypothese vorstellen und diskutieren.

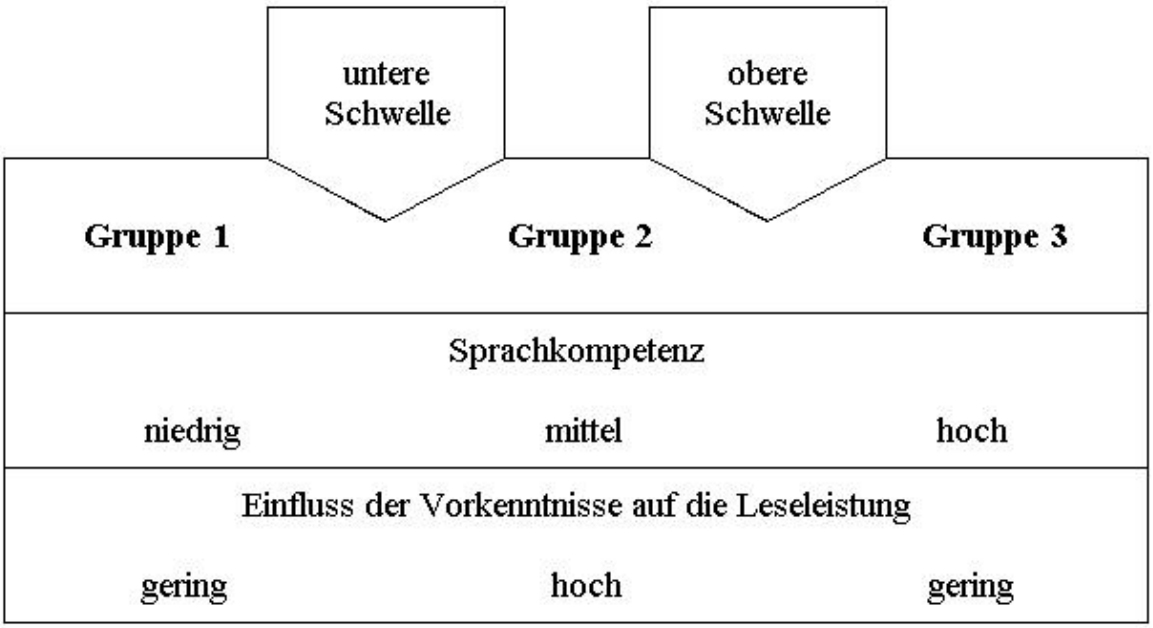


Abbildung 20: Fachkenntnisse, Sprachkenntnisse und Leseverstehen – die "Doppelte Schwellenhypothese"

Die Eckdaten der Studie **Claphams (1996; 2000)** wurden im Abschnitt "Studien zur Rolle der Vorkenntnisse" bereits vorgestellt. Hier geht es um die Beobachtungen zum Verhältnis von Sprachkompetenz in der Fremdsprache und Fachkompetenz beim Verständnis von Texten mit Fachbezug. Clapham stellte fest, dass der Einfluss der Fachkenntnisse auf das Textverständnis nicht konstant ist, sondern auch vom Niveau der Sprachkenntnisse abhängt. Vor allem bei einer mittleren Sprachkompetenz spielt das Hintergrundwissen eine große Rolle für das Textverständnis. Bei geringer Sprachkompetenz können Fachkenntnisse nicht über die sprachlichen Schwierigkeiten hinweghelfen. Bei hoher Sprachkompetenz können Wissenslücken durch den Einsatz der Sprachkompetenz ausgeglichen werden, so dass die Fachkompetenz nicht das bestimmende Element für das Verständnis von Fachtexten darstellt. Nur bei mittlerer Sprachkompetenz kommt den Fachkenntnissen also eine bedeutsame Rolle zu. Diesen Wirkungszusammenhang bezeichne ich als "Doppelte Schwellenhypothese" (siehe Abbildung 20). Allerdings können sich diese Zusammenhänge bei Texten mit einem besonders ausgeprägten Fachlichkeitsgrad auch verändern. Dann scheint die Bedeutung des Hintergrundwissens für das Verständnis zuzunehmen.

Zur Doppelten Schwellenhypothese gelangte Clapham über den Zusammenhang zwischen den Leistungen in einem Grammatiktest und den Leistungen im Test des Leseverstehens: Wenn Kandidaten im Grammatiktest weniger als 60 Prozent erzielten, schienen fortgeschrittene Fachkenntnisse keine positiven Auswirkungen auf das Verständnis von fachsprachlicher Literatur zu haben. Bei Kandidaten, die im Grammatiktest über 60 Prozent erzielten, hatten die Fachkenntnisse demgegenüber einen ausgeprägten Einfluss auf das Verständnis von fachsprachlicher Literatur. Kandidaten, die im Grammatiktest über 80 Prozent erzielten, verstanden Fachtexte unabhängig von dem Fachgebiet und unabhängig von ihren eigenen Fachkenntnissen besser als Kandidaten mit einem mittleren Ergebnis im Grammatiktest (unabhängig vom Niveau der Fachkenntnisse). Offensichtlich korreliert die Leistung, die im Grammatiktest gemessen wurde, positiv mit der Leistung, die beim Verständnis von Texten mit Fachbezug erforderlich ist. Dieser Zusammenhang konnte aber nicht auf jedem Sprachniveau beobachtet werden. Es schien vielmehr eine Schwelle zu geben, unter der die Kandidaten von ihren Fachkenntnissen nicht profitieren konnten. Nur bei Kandidaten mit einer mittleren Sprachkompetenz hatten die Fachkenntnisse starke Auswirkungen auf die Verstehensleistung.

Bei der Untersuchung bleiben einige Fragen offen:

- Warum wurde ein Grammatiktest als Indikator für "allgemeine Sprachkompetenz" gewählt? Wie in Kapitel 4.2 (Seite 114 ff) erläutert wurde, stellen Grammatiktests keine optimalen Tests für das Konstrukt "allgemeine Sprachfähigkeit" dar. Der Grammatiktest wurde offensichtlich nur gewählt, weil die Ergebnisse im Rahmen der IELTS-Durchführung vorlagen.

- Eine weitere offene Frage lautet: Warum wählte sie 60 und 80 Prozent als Schwellenwerte? Die Werte scheinen mehr oder weniger willkürlich gewählt worden zu sein; die Festlegung der kritischen Werte wird jedenfalls nicht begründet. Das arithmetische Mittel im Grammatiktest lag bei 69 Prozent mit einer Standardabweichung von 16,4 Prozent und dem Median bei 71 Prozent (Clapham, 1996: 105). Es hat den Anschein, als ob die Schwellenwerte mehr oder weniger willkürlich ungefähr zehn Prozent oberhalb und unterhalb des Mittelwerts angesetzt wurden. Allerdings zielte ihre Studie auch nicht darauf ab, die Schwellenwerte exakt zu bestimmen.

- Schließlich muss gefragt werden, ob die Ergebnisse aus Claphams Studie überhaupt als überzeugender Beleg für die Doppelte Schwellenhypothese angesehen werden können. Clapham verglich die Leseleistungen von Studierenden in Texten innerhalb und außerhalb ihres Fachgebiets. Nur bei einer der drei Konstellationen stimmten die Ergebnisse annähernd mit der Hypothese überein: Die Unterschiede zwischen den Ergebnissen von Studenten aus den Wirtschafts- und Sozialwissenschaften und von Studenten aus dem Bereich Biowissenschaften/Medizin waren nicht signifikant für Gruppen mit niedrigen Ergebnissen im Grammatiktest. Wurden Gruppen mit einem mittleren Ergebnis im Grammatiktest verglichen, ergaben sich auf dem Niveau 99 Prozent signifikante Unterschiede, bei Gruppen mit hohen Ergebnissen im Grammatiktest sank das Ergebnis auf 95 Prozent. Dies entspricht in etwa der Erwartung der Doppelten Schwellenhypothese. Claphams Studie umfasste zusätzlich Studenten aus technischen und naturwissenschaftlichen Disziplinen, doch die Ergebnisse mit dieser Studentengruppe wichen teilweise erheblich von dem erwarteten Muster ab. Clapham vermutete, dass einige Texte zu allgemein seien und daher das Ergebnis verfälschten. Auch als nur noch die Ergebnisse in Texten mit besonders hohem Fachlichkeitsgrad in die Analyse einbezogen wurden, gab es jedoch keine zusätz-

lichen Hinweise auf die Doppelte Schwellenhypothese.
Auch aus der Illustrationen der Daten in Streudiagrammen mit Differenzwerten sind keine überzeugenden Hinweise auf die Schwellenhypothese zu gewinnen. Clapham berechnet jeweils drei Anpassungslinien: für Ergebnisse bis 60 Prozent im Grammatiktest, zwischen 60 und 80 und über 80 Prozent. Nur in einem Fall folgen die Linien dem nach der Schwellenhypothese zu erwartenden Muster. (Streudiagramme mit Differenzwerten werden auch in meiner Studie eingesetzt; siehe Kapitel 6.2.3, Seite 310.)

Ebenfalls in der Mitte der 1990er Jahre führte **Ridgway (1997)** eine Studie zur Doppelten Schwellenhypothese durch. Seine Studie zielte ebenfalls eher darauf ab, die Schwellenhypothese zu bestätigen oder abzulehnen, nicht aber darauf, das genaue Sprachniveau der Schwellen zu lokalisieren. Er teilte 69 türkischen Studenten der Fachrichtungen Wirtschaft und Bauwesen in Gruppen mit hoher und niedriger Englischkompetenz ein. Das Niveau der Englischkompetenz wurde mit einem Lesetest erfasst. Zehn Studenten mit einem mittleren Ergebnis in diesem Test wurden nicht weiter berücksichtigt. Dann verglich Ridgway die Ergebnisse der Studenten in Leseverstehenstests zu einem wirtschaftlichen Thema und einem Thema aus dem Bereich des Bauingenieurwesens. Analog zur Doppelten Schwellenhypothese dürften weder Studenten mit niedrigen noch die Studenten mit weit fortgeschrittenen Sprachkenntnissen von Vorkenntnissen profitieren. Das heißt, zwischen den Leseleistungen unterschiedlicher (Fach-) Gruppen mit gleicher Englischkompetenz dürften keine großen Unterschiede auftreten. Sein Ergebnis wich von dieser Erwartung jedoch ab: Bei Studenten mit niedrigen Englischkenntnissen waren die Unterschiede zwischen den Ergebnissen in Leseverstehenstests aus dem eigenen Studienfach und den Ergebnissen in Tests aus dem fremden Studienfach tatsächlich nicht signifikant. Bei Studenten mit hohen Englischkenntnissen waren die Ergebnisse im Leseverstehen Bauwesen signifikant unterschiedlich, zum Wirtschaftstest jedoch nicht. Ridgways Schlussfolgerung lautet: "in some cases a background knowledge effect is detectable and in others it is not" (1997: 162). Die Ergebnisse seiner Studie können allenfalls als Bestätigung für die "untere Schwelle" interpretiert werden. Die "obere Schwelle" war nicht durchgehend nachzuweisen. Eine Ursache für die uneinheitlichen Ergebnisse liegt nach Ridgway in unterschiedlichen Verständlichkeitsgraden der Texte. Die kleine Probandenzahl dürfte ebenfalls einen Beitrag zu den wenig aussagekräftigen Ergebnissen geleistet haben.

Schlussfolgerungen

Die Studien zur Rolle der Vorkenntnisse bei Leseverstehenstests mit Fachbezug ergeben kein schlüssiges Bild. Beinahe alle Studien legen nahe, dass Vorkenntnisse eine positive Auswirkung auf das Leseverstehen haben. Der Einfluss der Variable Vorkenntnisse scheint jedoch geringer zu sein als der Einfluss der Fremdsprachenkenntnisse (Alderson, 1988; Alderson/Urquhart, 1985a; 1985b; Alvermann/Hynd, 1989; Birjandi/Alavi/Salmani-Nodoushan, 2002; Chen/Graves, 1995; Erickson/Molloy, 1983; Hock, 1990; Hung, 1990; Koh, 1985; Osman, 1984; Peretz/Shoham, 1990; Ridgway, 1997; Shoham/Peretz/Vorhaus, 1987; Tan, 1990; zum Hörverstehen: Jensen/Hansen, 1995; zum Sprechen: Douglas/Selinker, 1993; Papajohn, 1999; zum Schreiben: Read, 1990; Tedick, 1990). Die Ergebnisse der Studien sind jedoch nicht widerspruchsfrei. Die höchsten Ergebnisse wurden nicht immer von derjenigen Gruppe erzielt, von der man es (wegen der Vorkenntnisse) erwartet hätte. Folgende Gründe wurden in den Studien angeführt: Es gibt methodische Schwierigkeiten bei der Erfassung der Vorkenntnisse und daraus resultierende Schwierigkeiten, Gruppen von Teilnehmern mit vergleichbaren Vorkenntnissen zusammenzustellen (z. B. Alderson/Urquhart, 1985a). Es ist schwierig, die Textschwierigkeit zu messen (z. B. Alderson/Urquhart, 1985a; 1985b; Birjandi/Alavi/Salmani-Nodoushan, 2002; Clapham, 1996; Shoham/Peretz/Vorhaus, 1987). Auch die Items von Leseverstehenstests haben einen Einfluss darauf, in welcher Weise die Testteilnehmer Vorkenntnisse einsetzen können (z. B. Peretz/Shoham, 1990). Schließlich wurde festgestellt, dass sich der Einfluss der Vorkenntnisse in Abhängigkeit vom Niveau der Fremdsprachenkenntnisse verändert (Alvermann/Hynd, 1989; Clapham, 1996; Hung, 1990; Jensen/Hansen, 1995; Ridgway, 1997).

Die Studien zur Rolle der Vorkenntnisse und besonders zur Schwellenhypothese hatten Auswirkungen auf den Einsatz von Sprachtests mit Fachbezug. Clapham sieht die Konstruktvalidität von Leseverstehensaufgaben mit Fachbezug kompromittiert. Sie hält Sprachtests mit Fachbezug für nicht ausreichend valide, weil der Einfluss der Vorkenntnisse nicht für alle Gruppen gleich sei:

> If further studies support these findings, it will be clear that subject-specific tests are not equally valid for academic learners at different levels of English language proficiency (Clapham, 2000: 516).

Der Einsatz von Sprachtests für den Hochschulzugang mit Fachbezug ist ihrer Ansicht nach nicht gerechtfertigt, wobei sie auch auf das bereits erwähnte Argument der Testökonomie verweist:

> ... in the light of recent research it seems that this ESAP [English for Special Academic Purposes] testing, which adds hugely to testers' and administrators' labours, gives no perceptible advantage to most students indeed, because of the inevitable lack of equivalence of some supposedly equivalent subtests, might disadvantage some of them. The tests, although designed with the best of intentions, might not therefore be valid, and it seems likely, with hindsight, that large-scale ESAP testing has led test constructors up a blind alley (Clapham, 2000: 516).

Diese Argumentation setzte sich für Sprachtests für den Hochschulzugang durch. Eine konkrete Auswirkung von Claphams Untersuchung war eine Änderung von IELTS. Im Zuge einer Überarbeitung der Prüfung wurden die Lesetexte mit Fachbezug durch Texte ersetzt, welche keinen Bezug zu einem bestimmten Studienfach aufweisen (Clapham, 2000: 516). Sprachtests für den Beruf blieben von ihrer Studie weitgehend unberührt. Dort erfreuen sich Sprachtests mit Fachbezug weiterhin großer Beliebtheit. Ihre Zahl scheint im Gegenteil deutlich zuzunehmen. Über diese beiden Positionen denkt Douglas nach:

> A second direction which may continue in the future is a possible tendency toward less and less specificity in LSP [languages for specific purposes] tests. For example, we have seen that the ELTS/IELTS has progressively become less specific in subsequent revisions. [...] It may be, too, that test developers and researchers are less and less interested in what makes language use situations and texts different and more interested in what features they share. This trend is related to a concern for generalizability, the degree to which performance on a test can lead to inferences about performance in non-test situations (Douglas, 1997: 117-118).

Mit Blick auf Sprachtests für den Hochschulzugang und den Bereich der Sprachtests für das Studium regt Clapham an, Texte und Items zu wählen, die *aussehen* wie Sprachtests mit Fachbezug, die jedoch auch ohne einschlägige Fachkenntnisse bearbeitet werden können:

> The main point I make in this paper is that EAP [English for academic purposes] test candidates should not be penalised simply because they have not previously encountered Western type academic discourse practices. This means that future EAP tests should only *appear* to be testing academic discourse. They should appear to do so in order that EAP preparation classes will focus not only on individual aspects of the English language, but also on the genres and functions used in academic discourse (Clapham, 2000: 519; Hervorhebung im Original).

Auch Fulcher möchte nicht auf die vermeintlichen positiven Auswirkungen von Sprachtests mit Fachbezug verzichten und regt ebenfalls die Konzeption von Tests mit geringem Fachlichkeitsgrad an, welche wie Sprachtests mit Fachbezug wirken:

> On the one hand, test takers need to perceive the test as relevant to their subject and studies to achieve response validity. On the other, test content, title and labels of sub-tests may have significant washback effect upon what teachers do in classrooms. New tests may therefore continue to look very similar to their ancestors, but score meaning will be established in the light of construct validity

> studies rather than merely test content. However, unless future research (such as that into performance testing) can provide new and measurable definitions of 'specific', it may no longer be appropriate to talk about tests of English *for* Academic Purposes, but rather of tests of English *through* Academic Contexts (EAC); (Fulcher, 1999: 234-244; Hervorhebungen im Original).

Fulcher weist an dieser Stelle auf die Notwendigkeit hin, Argumente für eine hohe Konstruktvalidität zu suchen und bei Sprachtests mit Fachbezug, welche häufig als direkte Tests angesehen werden, nicht automatisch von einer hohen Konstruktvalidität auszugehen.

Aber wie sehen diese Tests aus? Man wird in der Praxis bei dem Vorhaben, Sprachtests für den Hochschulzugang nach den Empfehlungen von Clapham und Fulcher zu konzipieren, auf Schwierigkeiten treffen bzw. man benötigt weitere Informationen, etwa über das Niveau etwaiger sprachlicher Schwellen, bei denen sich der Einfluss der Vorkenntnisse ändert. Ähnlich offen ist die Frage nach der Textauswahl: Welcher Fachlichkeitsgrad ist im Sinne der obigen Argumentation förderlich?

Die Befunde zur "Doppelten Schwellenhypothese" überzeugen nicht in allen Punkten. Dennoch wird die Schwellenhypothese in keiner Studie explizit widerlegt. Die Schwellenhypothese könnte Testerstellern neue Perspektiven bieten, daher lohnt sich eine genaue Betrachtung. Die Schwellenhypothese fand bereits Eingang in Theorien zum Lesen und zum Testen. Alderson sieht ebenfalls die Notwendigkeit von weiteren Studien und hofft, dass man mit der Schwellenhypothese Aussagen über die Angemessenheit von Leseverstehenstests für bestimmte Gruppen von Testteilnehmern treffen kann, wenn die Textschwierigkeit einbezogen wird.

> Needless to say, Clapham's results need replication and extension. Nevertheless, they suggest that language testers might some day be able to define text difficulty in terms of what level of language abilities a reader must have in order to understand that particular text, and vice versa, what sort of text a learner of a given level of language ability might be expected to be able to read (Alderson, 2000a: 104).

Für mich war dieser Gedanke eine zentrale Motivation zur Durchführung des Forschungsprojekts, das ich in Kapitel 6 vorstelle. Es wäre für die Testerstellung nützlich, wenn man mit der Doppelten Schwellenhypothese den Einfluss der Vorkenntnisse bei einem Lesetext mit einem bestimmten Fachlichkeitsgrad für Gruppen mit einer bestimmten Sprachkompetenz vorhersagen könnte. Im Falle von Sprachtests für den Hochschulzugang wäre es sinnvoll und notwendig, unterschiedliche Gruppen von Kandidaten zu betrachten:

Kandidaten mit einer hohen Deutschkompetenz sollten in einem Sprachtest für den Hochschulzugang ein gutes Ergebnis erzielen. Dies scheint sowohl in Sprachtests mit Fachbezug als auch in Sprachtests ohne Fachbezug einzutreten, denn Kandidaten mit einer hohen Sprachkompetenz können Wissenslücken durch den Einsatz der umfassenden Sprachkenntnisse ausgleichen. Nur bei Texten mit besonders ausgeprägtem Fachlichkeitsgrad spielen auch bei Kandidaten mit einer hohen Deutschkompetenz die Fachkenntnisse eine zusätzliche Rolle.

Kandidaten mit einer geringen Deutschkompetenz sollten einen Sprachtest für den Hochschulzugang unabhängig von ihren Fachkenntnissen nicht bestehen. Nur bei einer Sprachkompetenz über einer gewissen Schwelle können Fachkenntnisse bei der Bewältigung von fachsprachlichen Aufgaben helfen. Es ist – so hat Claphams Studie gezeigt – unwahrscheinlich, dass jemand mit geringen Sprachkenntnissen Fragen zu einem fremdsprachigen Fachtext beantworten kann, obwohl er mit dem Inhalt des Fachtextes im Prinzip vertraut ist. Es besteht also kein Grund zur Annahme, dass Kandidaten mit geringen Deutschkenntnissen in einem Deutschtest mit Fachbezug allein wegen hoher Fachkenntnisse ein gutes Ergebnis erzielen, obwohl die Deutschkenntnisse für die Aufnahme eines Studiums nicht ausreichen.

Bei einem Text mit einem hohen Grad an Fachsprachlichkeit und einer Sprachkompetenz auf mittlerem Niveau spielt die Fachkompetenz laut Claphams Studie eine erhebliche Rolle für den Verstehensprozess. Für Sprachtests mit Fachbezug bedeutet dies, dass die Fachkompetenz die gezeigte sprachliche Leistung beeinflusst. Mit Blick auf Sprachtests für den Hochschulzugang ist zu fragen, wie groß diese Gruppe ist und über welchen Sprachstand die Kandidaten verfügen. Wenn der Sprachstand, bei dem die Vorkenntnisse beim Leseverstehen helfen, bereits auf einem so hohen Niveau ist, dass die Studienbewerber unabhängig von den Vorkenntnissen über ausreichende Sprachkenntnisse für die Aufnahme eines Fachstudiums verfügen, wäre der Einsatz von Sprachtests mit Fachbezug nicht unfair. Wenn die Kandidaten aus der Gruppe über einen so niedrigen Sprachstand verfügen, dass sie mit oder ohne Vorkenntnisse eine Sprachprüfung für den Hochschulzugang nicht bestehen würden, wäre der Einsatz von Sprachtests mit Fachbezug ebenfalls unproblematisch. Schwieriger wäre die dritte Möglichkeit: Wenn die Kandidaten mit "mittleren Sprachkenntnissen" eine Sprachprüfung für den Hochschulzugang nur dann bestehen, wenn sie ihre Vorkenntnisse

einsetzen, käme das Argument der fehlenden Testfairness in der Tat zum Tragen. Man könnte allerdings fragen, ob dieser Effekt in Einzelfällen nicht sogar wünschenswert ist. Warum sollen Kandidaten mit einer mittleren Deutschkompetenz ihre Fachkenntnisse, die ihnen auch im Studium helfen werden, nicht auch in einem Sprachtest einsetzen können? Wenn die Studienbewerber in der Lage sind, über ihr Hintergrundwissen sprachliche Lücken auszugleichen, verfügen sie über Fähigkeiten, die ihnen im Studium helfen können.

Die in diesem Kapitel vorgestellten Argumentationen werden im Kapitel 6 an einem konkreten Beispiel auf den Prüfstand gestellt.

6. Studie zum Fachbezug in Leseverstehenstests

Übersicht: Kapitel 6

In diesem Kapitel stelle ich eine Studie zu Leseverstehenstests mit Fachbezug vor. Diese Studie bezieht sich auf folgende Fragestellungen: Erstens, welche Rolle spielen Vorkenntnisse in Sprachtests mit Fachbezug? Zweitens, wie groß ist der Einfluss der Vorkenntnisse im Vergleich mit dem Einfluss der Sprachkenntnisse? Drittens, ändert sich die Intensität des Einflusses bei bestimmten sprachlichen Schwellen? Am Beginn des Kapitels steht ein Resümee der bisherigen Überlegungen zu Sprachtests mit Fachbezug, aus dem die Fragestellungen entwickelt werden.

Die Bedenken gegen Sprachtests mit Fachbezug, die ich in Kapitel 5.1 vorstellte, und die Ergebnisse der Studien, die in Kapitel 5.2 zusammengefasst wurden, führte bei den meisten Sprachtests für das Studium zu einem Verzicht auf einen Fachbezug. Ich wies bereits darauf hin, dass Sprachtests "mit akademischem Inhalt, aber ohne Fachbezug" jedoch ebenfalls mit Schwierigkeiten behaftet sind und nicht notwendigerweise eine zufrieden stellende Lösung des Problems darstellen. Hinzu kommt, dass ich der Vermittlung von Fachsprache auch in der Studienvorbereitung eine große Bedeutung beimesse. Lediglich eine Konzentration auf allgemeine Lern- und Arbeitstechniken wie von Jordan (1997) propagiert und eine daraus resultierende Beliebigkeit der Inhalte bilden meiner Ansicht nach kein tragfähiges Konzept für eine sprachliche Studienvorbereitung. Eine Förderung der Sprachkompetenz in der Studienvorbereitung, die

zusammen mit der Erweiterung der Fachkompetenz geschieht, ist zielgruppenadäquat und dem Ziel dienlich, sprachlich auf ein Fachstudium vorzubereiten.

Es sind mithin vor allem drei Gründe, die mich bewogen, mich auf das knifflige Gebiet der Sprachtests mit Fachbezug einzulassen: die erwarteten, positiven Auswirkungen von Sprachtests mit Fachbezug auf die sprachliche Studienvorbereitung, die Schwierigkeiten mit der Alternative, Sprachtests ohne Fachbezug, sowie die Erwartung, dass sich Zusammenhänge zwischen Vorkenntnissen und Sprachkenntnissen produktiv einsetzen lassen.

Die nachfolgende Studie beruht auf folgenden Annahmen: Wenn die Doppelte Schwellenhypothese zutrifft, sollte man mehr über die Schwellen und die Einflussgrößen erfahren. Möglicherweise können die Informationen in Sprachtests für den Hochschulzugang produktiv genutzt werden. Welche Gruppe von Studienbewerbern kann ihre Vorkenntnisse besonders einsetzen? Möglicherweise reichen die Sprachkenntnisse der Kandidaten aus der Gruppe "mit mittlerer Sprachkompetenz" ohnehin nicht für die Aufnahme eines Fachstudiums aus, möglicherweise würden sie einen Sprachtest für den Hochschulzugang unabhängig von den Vorkenntnissen nicht bestehen. Es ist also nötig, weitere Erfahrungen mit den Schwellenwerten zu sammeln, damit ein genaueres Bild über die Folgen eines Fachbezugs in einem Sprachtest für den Hochschulzugang für die Zulassungsentscheidung entstehen kann.

Die Aussagekraft der bestehenden Studien zur "Doppelten Schwellenhypothese" leiden unter methodischen Problemen. Diese sind in weiteren Studien zu verhindern:

- Erstens, um sprachliche Schwellen zu beschreiben, müssen jeweils mehrere mögliche Schwellenwerte betrachtet werden. Dass sprachliche Schwellen völlig unabhängig etwa vom Grad der Textschwierigkeit stets auf dem gleichen Niveau liegen, ist keine sinnvolle Annahme.

- Zweitens können sprachliche Schwellen nur mit Bezug zu einer reliablen Messgröße der Fremdsprachenkompetenz erfasst und beschrieben werden. Die Hypothese bezieht sich auf die Fremdsprachenkompetenz; es bedarf daher geeigneter Testverfahren für dieses Konstrukt. Ein fachsprachlicher Lesetext (wie bei Ridgway, 1997) oder ein Grammatiktest (wie bei Clapham, 1996) sind nicht geeignet.

- Drittens spielt der Fachlichkeitsgrad der Lesetexte eine Rolle. Will man Leistungen in fachsprachlichen Lesetests vergleichen, muss sich der Fachlichkeitsgrad der eingesetzten Texte ähneln.

6.1. Fragestellung und Methode

Übersicht: Kapitel 6.1

Am Anfang des Kapitels werden die drei Leitfragen der Studie vorgestellt. Es folgt eine Beschreibung der Methode: Vorgestellt und begründet werden die Auswahl der Texte und Items und die Verfahren zur Erhebung der Deutschkenntnisse sowie der Vorkenntnisse. Außerdem werden die Testteilnehmer beschrieben und Hypothesen geäußert.

Die Fragestellungen der Studie lauten:

Fragestellung 1

Erzielen ausländische Studienbewerber bessere Ergebnisse in Leseverstehenstests mit Fachbezug, wenn sie über Vorkenntnisse verfügen?

Fragestellung 2

Spielen die Fremdsprachkenntnisse oder die etwaigen Vorkenntnisse zum Thema die größere Rolle für die Ergebnisse in fremdsprachlichen Leseverstehenstests mit Fachbezug?

Fragestellung 3

Hängt der Einfluss der Vorkenntnisse vom Niveau der Deutschkenntnisse ab? Lassen sich sprachliche Schwellen ermitteln und beschreiben, bei denen sich der Einfluss der Vorkenntnisse zum Thema verändert?

In dieser Studie sollen am Rande auch methodische Aspekte berücksichtigt werden. Die drei Methoden zur Erhebung der Vorkenntnisse werden auf den Prüfstand gestellt: Benötigt man mehrere Erhebungen der Vorkenntnisse? Erfassen die unterschiedlichen Erhebungsformen tatsächlich die Vorkenntnisse?

Insgesamt wurden sechs Variablen erhoben:

- INFLATION: Ergebnisse im Leseverstehenstest mit Bezug zu einem wirtschaftlichen Thema "Inflation". Ziel: Erfassung der fachsprachlichen Lesekompetenz im Fach Wirtschaft.
- GESCHWINDIGKEIT: Ergebnisse im Leseverstehenstest mit Bezug zu einem technischen Thema "Geschwindigkeit". Ziel: Erfassung der fachsprachlichen Lesekompetenz zu einem technischen Thema.
- C-TEST: Ergebnisse in einem C-Test. Ziel: Erfassung der allgemeinen Sprachkompetenz
- BEGRIFFE: Kenntnis der Schlüsselbegriffe aus den Texten "Inflation" bzw. "Geschwindigkeit" vor dem Lesen. Ziel: Erfassung der Vorkenntnisse.
- BEKANNT: Vertrautheit mit dem Thema der Texte "Inflation" bzw. "Geschwindigkeit" nach dem Lesen. Ziel: Erfassung der Vorkenntnisse.
- KURS: Studienwunsch bzw. Kurszuordnung im Studienkolleg. Ziel: Erfassung der Vorkenntnisse bzw. Erfassung der Ergebnisse nach den Gruppen, für die die Tests in der Praxis erstellt würden.

Für die Studie stellte ich ein Testpaket zusammen, das aus den folgenden Elementen bestand:

- Fragebogen zur Person und zu Vorkenntnissen
- C-Tests
- Leseverstehenstest "Inflation"
- Leseverstehenstest "Geschwindigkeit"
- Frage zur Vorkenntnissen

Das Testpaket wurden über 500 Kandidaten aus verschiedenen Studienkollegs vorgelegt. Ungefähr 150 Datenerhebungen wurden von mir durchgeführt, 350 erfolgten durch Kolleginnen und Kollegen anderer Studienkollegs, wobei ich die Korrektur und Auswertung vornahm. Die Auswahl der Studienkollegs hing von der Bereitschaft zur Mitarbeit ab. Zur Bearbeitung erhielten die Kandidaten 90 Minuten Zeit. Diese waren wie folgt aufgeteilt: 10 Minuten Einführung, 20 Minuten Fragebogen und C-Tests, 30 Minuten für jeden Leseverstehenstest.

6.1.1. Probanden

An den Tests "Inflation" und "Geschwindigkeit" nahmen Kandidaten von mehreren Studienkollegs teil. Dazu gehörten Kollegiaten des ersten und zweiten Semesters sowie aus unterschiedlichen Kursen. Insgesamt nahmen über 500 Kandidaten aus acht Studienkollegs teil. Bei den Kollegiaten handelt es sich um ausländische Studienbewerber, die aufgrund ihrer bisherigen Qualifikationen aus dem Heimatland keine Direktzulassung an einer deutschen Hochschule erhalten können. Sie wurden im Studienkolleg einem studienvorbereitenden Kurs zugeordnet, welcher je nach Studienziel auf ein Studium bestimmter Fächer vorbereitet. Im Rahmen dieser Kurse erhalten sie nicht nur Deutschunterricht, sondern je nach Kurs auch Fächer, die für das Studienziel relevant sind.

Die Stichprobe war in Bezug auf die Herkunftssprachen bzw. Herkunftsländer durchaus repräsentativ für ausländische Studierende an deutschen Hochschulen (vgl. die Angaben zu der Zusammensetzung ausländischer Studierender 2003 in DAAD/HIS, 2004: 14-17). Die Gruppe der Chinesinnen und Chinesen war aber deutlich überrepräsentiert. Die größte Gruppe der 486 Studierenden, die beide Leseverstehenstests bearbeiteten, stammte aus China (n = 106; 21.8 %), gefolgt von Studierenden aus arabischen Ländern (n = 94; 19.3 %) bzw. aus Russland und Mittel- und Osteuropa (n = 89; 18.3 %). Weitere nennenswerte Gruppen kamen aus Südostasien (Vietnam, Indonesien, Thailand: n = 51; 10.5 %), aus Mittel- und Südafrika (n = 45; 9.3 %) und aus Südamerika (n = 28; 5.8 %). Die sprachlichen Leistungen der Studentengruppen waren unterschiedlich (siehe Tabelle 28). Während die chinesischen Studierenden in allen drei Tests (INFLATION, GESCHWINDIGKEIT, C-TEST) unterdurchschnittlich abschnitten, lagen die Leistungen der Studierenden aus Russland und Mittel- und Osteuropa über dem Durchschnitt. Studierende aus arabischen Ländern erzielten etwa durchschnittliche Leistungen. Die Teststärke (Größe der Stichprobe) reichte nicht aus, um zu erheben, ob der Einfluss der Vorkenntnisse abhängig von der Herkunft unterschiedlich stark war. Die meisten Studienbewerber aus arabischen Ländern strebten beispielsweise ein technisches Studienfach an; es gab nur 9 Studienbewerber mit dem Studienziel Wirtschaft aus arabischen Ländern. Es ist also zu vermuten, dass der achtprozentige Unterschied

zwischen den Leistungen in INFLATION und GESCHWINDIGKEIT nicht länderabhängig ist, sondern sich aus unterschiedlichen Vorkenntnissen erklärt.

Tabelle 28: Leistungen der Kollegiaten in den Tests der Studie (Gruppen nach Herkunftsländern)

Variable	Studierende aus China	Studierende aus arabischen Ländern	Studierende aus Russland und Mittel- und Osteuropa	Gesamte Stichprobe
INFLATION	46,0 %; n = 106	40,1 %; n = 94	64,3 %; n = 89	51,5 %; n = 486
GESCHWINDIGKEIT	43,1 %; n = 106	48,0 %; n = 94	63,3 %; n = 89	52,8 %; n = 486
C-TEST	39,8 %; n = 105	57,1 %; n = 92	67,7 %; n = 88	55,6 %, n = 480

6.1.2. Die Leseverstehenstests mit Fachbezug

Die Texte der Leseverstehenstests

Die Auswahl der Texte ist ein konstitutives Merkmal der Argumentation. Es sollten Texte gewählt werden, die sich für einen Einsatz in einem Sprachtests mit Fachbezug für den Hochschulzugang eignen. Dies sollte durch folgende Kriterien erfüllt werden:

- ähnlicher sprachlicher und fachlicher Schwierigkeitsgrad,
- deutlich erkennbarer Fachbezug,
- eher niedriger Fachlichkeitsgrad durch Erläuterung der Schlüsselbegriffe.

Die Wahl fiel auf die Texte "Inflation" (siehe Abbildung 21, Seite 238)und "Geschwindigkeit" (siehe Abbildung 22, Seite 238), welche aus dem Duden-Schülerlexikon stammen. Das mehrbändige Schülerlexikon, richtet sich an Schüler, es enthält laut Klappentext den Stoff bis zum Abitur. Die einzelnen Texte unterscheiden sich zwar sprachlich voneinander, denn es sind verschiedene Autoren am Werk und die einzelnen Disziplinen und Themen verlangen unterschiedliche Kommunikationsformen. Diese Reihe wird aber von einem Redaktionsteam herausgegeben, das offensichtlich mit Blick auf Sprache, Inhalt, Vorgehensweise und Lay-out für alle Bände einheitliche Kriterien anlegt. Dies schienen gute Voraussetzungen für die Erfüllung des Kriteriums "ähnlicher sprachlicher Schwierigkeitsgrad" zu sein. Die authentischen Texte aus der Studienvorbereitung müssten von deutschen Abiturienten unabhängig vom Studienziel und von den gewählten Kursen in der Oberstufe als Nachschlagewerk konsultiert werden können. Die Texte sind nicht journalistisch oder populärwissenschaftlich aufbereitet. Im Mittelpunkt stehen die Sachinformationen.

Auch im Lay-out sollten sich die Texte ähneln. Beide Texte enthielten eine Abbildung und eine Formel. Hier könnte man argumentieren, dass die Verwendung von Abbildungen dem eigentlichen Testkonstrukt von Leseverstehenstests widerspricht. Viele Texte, die im Studium eingesetzt werden, enthalten jedoch ebenfalls Abbildungen, und der Umgang damit stellt einen Teil des Testkonstrukts von Leseverstehenstests dar.

Um mich über den Fachlichkeitsgrad der Texte "Inflation" und "Geschwindigkeit" zu vergewissern, wählte ich Vergleichstexte. Die Vergleichstexte "Inflation" (Abbildung 23, Seite 239) und "Radar" (Abbildung 24, Seite Seite 239) haben ähnliche Themen. Sie

sind aber für eine andere Zielgruppe verfasst worden und verfügen über einen ausgeprägten Fachlichkeitsgrad. Sie richten sich weniger an junge Erwachsene vor der Aufnahme eines Studiums, als vielmehr an Studierende und Fachleute. Durch einen Vergleich mit diesen Texten sollte der Fachlichkeitsgrad der Texte "Inflation" und "Geschwindigkeit" weiter bestimmt werden.

Geschwindigkeitsmessung

Die Geschwindigkeit eines Körpers kann allgemein berechnet werden mit den Gleichungen

$$v = \frac{s}{t} \quad \text{oder} \quad \Delta v = \frac{\Delta s}{\Delta t}$$

Dabei bedeuten: v Geschwindigkeit, s zurückgelegter Weg, t benötige Zeit.

Bei der Geschwindigkeit ist zwischen der Durchschnittsgeschwindigkeit und der Augenblicksgeschwindigkeit (Momentangeschwindigkeit) zu unterscheiden. Die Durchschnittsgeschwindigkeit gibt an, wie groß die mittlere Geschwindigkeit längs einer Strecke ist, die ein Körper in einer bestimmten Zeit zurücklegt. Die Augenblicksgeschwindigkeit gibt die Geschwindigkeit zu einem bestimmten Zeitpunkt an. Je nach der Art der Messung erhält man entweder Durchschnittsgeschwindigkeiten oder näherungsweise Augenblicksgeschwindigkeiten. Die Zeitmessung kann mit einer Stoppuhr, aber auch elektrisch erfolgen.

Quelle: Basiswissen Schule-Physik, 2005.

(vollständiger Text: Anhang 7, Seite 364)

Abbildung 21: Text "Geschwindigkeit"

Inflation und Deflation

Inflation ist die über einen längeren Zeitraum festzustellende Erhöhung des Preisniveaus. Dabei ist die Zunahme der Preise im Durchschnitt, bei Beachtung der Gewichtung der Waren, von Bedeutung und nicht die Erhöhung einzelner Preise. Man misst die Inflation mit dem Preisindex für die Lebenshaltung aller privaten Haushalte. Die Ermittlung des Preisindexes erfolgt durch Feststellung der Entwicklung der Ausgaben des Durchschnittshaushaltes für einen repräsentativen Warenkorb einer Periode. Beispiel: Ausgaben für einen Warenkorb im Jahre x: 12000,-- DM; Ausgaben für einen Warenkorb im Jahre y: 12300,-- DM. Der Preisindex beträgt demnach:

$$\frac{12300}{12000} \times 100 = 102{,}5 = 2{,}5\% \text{ Inflationsrate}$$

Quelle: Basiswissen Schule-Wirtschaft, 2005.

(vollständiger Text: Anhang 10, Seite 370)

Abbildung 22: Text "Inflation

Inflation

I. Begriff: Prozess anhaltender Preisniveausteigerungen, die über eine gewissen Marge hinausgehen. Inflation ist nur als dynamischer Vorgang denkbar, bei dem Inflation aus einem bestimmten Ursachenkomplex im ökonomischen System entsteht und wieder auf dieses zurückwirkt. Zur Inflation zählen nur Steigerungen des Preisniveaus. Jene sind von Steigerungen der Einzelpreise zu unterscheiden, die zu den für eine Marktwirtschaft normalen Vorgängen zählen. Die Flexibilität der Einzelpreise hat für den Marktmechanismus die wichtige Funktion, die Produktionsfaktoren so zu lenken bzw. umzulenken, dass das Güterangebot dem Bedarf angepasst wird. Einzelpreissteigerungen (-senkungen) signalisieren den Anbietern c. p. einen höheren (geringeren) Bedarf, spiegeln also die relativen Knappheitsverhältnisse wider. Bei Preisnivaustabilität sind diese anhand der absoluten Preisänderungen unschwer zu erkennen. [...]

Quelle: Gabler Wirtschaftslexikon. 1997. 14. Auflage. Wiesbaden: Gabler, 1857-1863.

(vollständiger Text: Anhang 12, Seite 373)

Abbildung 23: Vergleichstext mit ausgeprägtem Fachlichkeitsgrad "Inflation"

Radar, *Radio Detection and Ranging*, Verfahren zur Entdeckung und Positionsbestimmung von festen und bewegten Objekten mit Hilfe elektromagnetischer Wellen.

Das Radar arbeitet nach dem Prinzip eines Echolots: Der Radarsender strahlt elektromagnetische Wellen im mm- bis m-Bereich aus, deren Reflexionen ausgewertet werden. Der Ort eines vom Radar erfassten Objekts wird aus der Laufzeit und der Richtung des Echos bestimmt; unter Ausnutzung des Doppler-Effektes kann die Relativgeschwindigkeit zwischen Radargerät und Zielobjekt berechnet werden. Gegenüber optischen oder akustischen Ortungsverfahren besteht der Vorteil der Radartechnik im hohen Durchdringungsvermögen der Funkwellen und ihrer größeren Reichweite.

Beim Impulsradar werden die Funkwellen in Form kurzer Impulse (0,05-1μs) abgestrahlt. Die Vorteile dieses Betriebsregimes sind neben einer Energieersparnis die einfache Bestimmung der Laufzeit der Impulse und die Möglichkeit der Doppelnutzung der Radarantenne zum Senden und Empfangen. Die im Muttergenerator erzeugten Impulse werden gleichzeitig über den Modulator an den Sender und als Steuerimpuls an das Sichtgerät (bzw. die Auswerteelektronik) gegeben. Der Duplexer (Sende-Empfangs-Weiche) verhindert ein Übersprechen der Suchimpulse auf den Empfänger. Wird als Sichtgerät eine Elektronenstrahlröhre verwendet, so steuert der Muttergenerator die Zeitablenkung und der im Empfänger aufbereitete Echoimpuls die Vertikalablenkung oder die Intensität des Elektronenstrahls. Auf dem Bildschirm erscheint ein Zacken oder Leuchtfleck, dessen Lage durch die Laufzeit des Impulses bestimmt ist und somit der Entfernung zum reflektierenden Objekt entspricht. [...]

Quelle: *Lexikon der Physik: in sechs Bänden.* 1999. Band 4. Heidelberg: Spektrum Akademischer Verlag, 395-396.

(vollständiger Text: Anhang 9, Seite 368)

Abbildung 24: Vergleichstext mit ausgeprägtem Fachlichkeitsgrad "Radar"

Ob der Schwierigkeitsgrad der Leseverstehenstests "Inflation" und "Geschwindigkeit" tatsächlich vergleichbar ist und wie ausgeprägt ihr Fachlichkeitsgrad ist, soll im Folgenden diskutiert werden. In der Lesbarkeitsforschung sind quantitative und qualitative Methoden zur Erhebung des sprachlichen Schwierigkeitsgrads entwickelt worden. Die Einschätzung der Textschwierigkeit von Experten stellt einen qualitativen Ansatz dar, quantitativ wird Textschwierigkeit mit Lesbarkeits- bzw. Verständlichkeitsformeln erhoben (Ballstaedt/Mandl/Tergan, 1982; Bayer/Seidel, 1979; Biere, 1991; Briest, 1974; Flesch, 1949/1974; 1979; Fulcher, 1997; Grabowski, 1991; Groeben, 1982; Lutjeharms, 1988; Teigeler, 1979; Wetzchewald, o. J). Ich verwende folgende Verfahren:

- Erstens das "Hamburger Verständlichkeitskonzept",
- zweitens eine Beschreibung der Textschwierigkeit nach linguistischen Kriterien,
- drittens die Erhebung der Textschwierigkeit mit dem *Flesch-Index* und dem *Gunning-Fog-Index*.

Textschwierigkeit nach dem "Hamburger Verständlichkeitskonzept": Das von Langer, Schulz von Thun und Tausch entwickelte Verständlichkeitskonzept zur qualitativen Ermittlung der Textschwierigkeit basiert auf Expertenratings (Langer/Schulz von Thun/Tausch, 1999). Sie regen an, die Verständlichkeit von Texten nach folgenden Kriterien zu beurteilen: "Einfachheit", "Gliederung-Ordnung", "Kürze-Prägnanz" und "Anregende Zusätze". Im Rahmen der Arbeit möchte ich auf eine Widergabe der Erläuterung dieser Kriterien verzichten. Die Texte werden auf einer fünfstufigen Ratingskala bewertet:

- -	Merkmal nicht ausgeprägt
-	Merkmal wenig ausgeprägt
0	Merkmal weder stark noch schwach ausgeprägt
+	Merkmal stark ausgeprägt
+ +	Merkmal sehr stark ausgeprägt

Hinzuweisen ist noch auf folgenden Zusammenhang: Die Einschätzung "+ +" ist nicht unbedingt grundsätzlich ein Hinweis auf Verständlichkeit. Beim Merkmal Kürze-Prägnanz geht man beispielsweise davon aus, dass verständliche Texte weder zu kurz noch zu weitschweifig sein sollten. Auch konkurrieren Merkmale miteinander: Ein Text, der sehr viele anregende Zusätze enthält, wird weitschweifig, verstößt also gegen das Merkmal Kürze-Prägnanz.

Kritik gegen das "Hamburger Verständlichkeitskonzept" richtet sich gegen die Betrachtung der Verständlichkeit als textimmanentes Merkmal, welche den Verstehensprozess auf Seiten des Lesers nicht berücksichtigt (Ballstaedt/Mandl/Tergan, 1982; Groeben, 1982; Teigeler, 1979). Mit Blick auf das Untersuchungsdesign der vorliegenden Studie geht es jedoch gerade um die Verständlichkeit als Merkmal des Textes.

Meine Einschätzung der Verständlichkeit geht aus Tabelle 29 (Seite 241) hervor. Sie beruht auf der Beschreibung der Kriterien von Langer, Schulz von Thun und Tausch (Langer/Schulz von Thun/Tausch, 1999). Meiner Bewertung nach sind die Vergleichstexte deutlich schwieriger als die Texte, welche in den Leseverstehenstests verwendet werden. Die beiden Texte für die Leseverstehenstests sind meiner Einschätzung nach ungefähr gleich verständlich.

Tabelle 29: Lesetexte mit Fachbezug –Textverständlichkeit nach dem "Hamburger Verständlichkeitskonzept"

Inflation		**Vergleichstext: Inflation**	
Einfachheit: 0	Gliederung-Ordnung: +	Einfachheit: - -	Gliederung-Ordnung: 0
Kürze-Prägnanz: +	Anregende Zusätze: 0	Kürze-Prägnanz: +	Anregende Zusätze: - -
Geschwindigkeitsmessung		**Vergleichstext: Radar**	
Einfachheit: -	Gliederung-Ordnung: +	Einfachheit: - -	Gliederung-Ordnung: 0
Kürze-Prägnanz: +	Anregende Zusätze: 0	Kürze-Prägnanz: +	Anregende Zusätze: -

Beschreibung der Textschwierigkeit nach linguistischen Kriterien: Auch eine linguistische Beschreibung der Texte führt zu einer vergleichbaren Einschätzung: In welcher Häufigkeit werden typische Textsortenmerkmale von Fachsprache verwendet? Fachsprachliche Merkmale können für unterschiedliche sprachliche Ebenen in Abhängigkeit von unterschiedlichen Textsorten beschrieben werden (z. B. Buhlmann/Fearns, 2000: 15-80; Hofmann, 1985: 72-242). Anhand einer Auswahl typischer, sprachlicher Merkmale von Texten mit hohem Fachlichkeitsgrad soll exemplarisch an den Texten "Inflation" und "Geschwindigkeit" sowie zum Vergleich an den Vergleichstexten "Inflation" und "Radar" der Grad der Fachlichkeit beschrieben werden (siehe auch Tabelle 30, Seite 243).

Die häufige Verwendung von Sätzen mit Passiv ist ein Merkmal, das für die Bestimmung des Fachlichkeitsgrad von Texten herangezogen werden kann. Im Text "Inflation" enthalten 29 Prozent aller Sätze eine Passivkonstruktion, im "Vergleichstext Inflation" sind es 40 Prozent. In den anderen Texten ist der Anteil deutlich höher: Bei "Geschwindigkeit" sind es 53 Prozent, beim "Vergleichstext Radar" 57 Prozent. Hier könnte das Thema und unterschiedliche Textsorten dazu geführt haben, dass die Texte zum Wirtschaftsthema weniger passivlastig sind als die Texte zum Technikthema. In den Techniktexten werden Prozesse beschrieben, in den Wirtschaftstexten geht es um Definitionen sowie die Darstellung von Ursachen und Folgen.

Typisch für Fachtexte mit hohem Fachlichkeitsgrad ist die häufige Verwendung von Genitiven zur Präzisierung. Der Anteil der Sätze mit Genitiv liegt bei "Inflation" und "Geschwindigkeit" mit 46 bzw. 47 Prozent deutlich unter demjenigen der Vergleichstexte ("Vergleichstext Inflation": 65 %, "Vergleichstext Radar": 81 %).

Unterschiedlich häufig werden Partizipialkonstruktionen als Attribut verwendet. In "Inflation" enthalten 21 Prozent der Sätze ein Partizipialattribut, in "Geschwindigkeit" sind es nur 13 Prozent. Bei den Vergleichstexten sind deutlich mehr Sätze mit Partizipialattribut anzutreffen: 40 Prozent im "Vergleichstext Inflation" bzw. 57 Prozent im "Vergleichstext Radar".

In allen vier Texten trifft man eine häufige Verwendung von Wortzusammensetzungen an. In den Vergleichstexten spiegelt sich in den Wortzusammensetzungen jedoch die größere inhaltliche Differenziertheit. "Inflation" enthält die folgenden Komposita mit

"Preis-": "-niveau", "-index", "-steigerung", "-erhöhung", "-anstieg" sowie "Preissteigerungsrate", dem einzigen Begriff mit drei Bestandteilen. Im "Vergleichstext Inflation" ist die Zahl höher, die Begriffe sind differenzierter und schwieriger zu verstehen: "Preisniveausteigerung", "Einzelpreissteigerung", "Preisniveaustabilität", "Preisbewegungen", "Preis- und Einkommenspolitik" usw. Ähnlich verhält es sich mit den "Geschwindigkeit" und dem Vergleichstext "Radar".

Weitere Unterscheidungen aufgrund fachsprachlicher Merkmale fallen nicht auf oder sind wegen der geringen Textumfänge nicht nachweisbar: Bei allen Texten ist eine geringe Bedeutung der Verben ("Deverbalisierung") festzustellen. Funktionsverben und Nominalisierungen kommen in allen Texten vor ("Letztlich *kommt* das in einer sinkenden Kaufkraft der Währungen *zum Ausdruck*." – Inflation; "die Möglichkeit der *Doppelnutzung* der Radarantenne zum *Senden* und *Empfangen*." – Vergleichstext Radar). Mit Ausnahme von "Inflation" sind alle Texte durchgängig im Präsens verfasst. Es werden ausschließlich die 3. Person Singular und Plural verwendet. Der Text "Geschwindigkeit" sowie die Vergleichstexte "Inflation" und "Radar" enthalten Konditionalsätze ohne Einleitungswort. Da die Anzahl der Konditionalsätze maximal drei beträgt, taugt dieses Merkmal ebenfalls nicht, um unterschiedliche Fachlichkeitsgrade zu illustrieren, welche sich in der Sprache äußern.

Tabelle 30: Lesetexte mit Fachbezug – einige fachsprachliche Merkmale (Häufigkeiten)

sprachliches Merkmal	**Inflation**	**Geschwindigkeit**	**Vergleichstext Inflation**	**Vergleichstext Radar**
Sätze mit Passiv (Anteil in %)	29 %	53 %	40 %	57 %
Sätze mit Genitivkonstruktionen (Anteil in %)	46 %	47 %	65 %	81 %
Sätze mit Partizipialattributen (Anteil in %)	21 %	13 %	40 %	57 %
Wortzusammensetzungen	häufig	häufig	sehr häufig, sehr differenziert	sehr häufig, sehr differenziert

Insgesamt unterstützt die sprachliche Analyse die Einschätzung der Textschwierigkeit nach dem Verständlichkeitskonzept: Die Texte aus dem Duden-Schülerlexikon verfügen über einen wahrnehmbaren Fachbezug, der sich auch an sprachlichen Merkmalen äußert. Er ist jedoch im Vergleich mit ausgewiesenen Fachtexten deutlich geringer ausgeprägt.

Ermittlung der Textschwierigkeit mit Indices: Zur Ermittlung des Schwierigkeitsgrads von Lesetexten gibt es eine Reihe von *quantitativen* Verfahren, die zunächst mit Blick auf das Leseverstehen in der Muttersprache und als Orientierungshilfe bei der Vereinfachung von Texten entwickelt wurden (Ballstaedt/Mandl/Tergan, 1982; Bayer/Seidel, 1979; Biere, 1991; Briest, 1974; Flesch, 1949/1974; 1979; Grabowski, 1991; Groeben, 1982; Mihm, 1973; Teigeler, 1979). Sie beziehen sich auf oberflächensprachliche Aspekte wie Satz- oder Wortlänge. Eine hohe durchschnittliche Satzlänge wird beispielsweise als Hinweis auf syntaktische Komplexität interpretiert oder eine hohe durchschnittliche Wortlänge als Hinweis auf die Schwierigkeit der Lexik. Zumindest eine große durchschnittliche Wortlänge ist ein häufiges Merkmal fachsprachlicher Texten (Hoffmann, 1985: 135-136 u. 204-205). In der Praxis dürften sich jedoch viele leicht verständliche Texte mit langen Sätzen finden lassen. Schwer verständlich werden Sätze durch die Komplexität der syntaktischen Struktur, welche durch die Satzlänge nicht in jedem Fall abgebildet wird (Grotjahn, 2000b: 26).

Die Kritik an derartigen Indices lautet, dass durch die Erfassung von oberflächlichen Textmerkmalen die Textverständlichkeit nicht umfassend beschrieben werden könne. Dieser Kritikpunkt war übrigens ein Ansatzpunkt für die Entwicklung des oben erwähnten "Hamburger Verständlichkeitskonzepts". In Diskussionen um derartige Indexzahlen wird außerdem hervorgehoben, dass weder Motivation und Voraussetzungen des Lesers noch die große Zahl weiterer Textmerkmale (z. B. Kohärenz, Gliederung) berücksichtigt werden (Carrell, 1987: 25; Nebe, 1990: 351). Im Zusammenhang mit der Schwierigkeit von Texten in Sprachtests spielen diese Aspekte jedoch keine entscheidende Rolle. Die Formeln zur Bewertung der Textverständlichkeit bieten hier eine Orientierungshilfe für die Texterstellung bzw. -auswahl. Problematisch ist bei Sprachtests für den Hochschulzugang allerdings die Übertragung auf das Verstehen in der Fremdsprache. Dies wird deutlich, wenn die Sprache in Comics betrachtet wird: Sie erweist sich als "sehr leicht" oder "sehr verständlich", wenn man sie mit einer Formel

zur Erfassung der Textverständlichkeit misst. Das trifft auf die Muttersprache möglicherweise zu, in der Fremdsprache ist das Verstehen der Sprechblasen in Bildgeschichten häufig aber anspruchsvoll.

Ich stelle zwei Indices vor, die sehr verbreitet sind. Der *Flesch Readability Index* bestimmt die Textschwierigkeit über die Länge der Wörter und die Länge der Sätze (Alderson, 2000a: 71; Flesch, o. J; 1949/1974; 1979.). Die Formel lautet:

Flesch Index = 206,835 – 84,6*Silben/Wörter – 1,015*Wörter/Sätze

Der Flesch Index liegt normalerweise auf einer Skala zwischen null und 100, wobei hohe Werte auf einen einfachen, leicht verständlichen Text hindeuten und niedrige Werte auf anspruchsvolle, schwer verständliche. Flesch stellt sogar einen direkten Zusammenhang zwischen Schulstufen und dem Flesch Index her (Flesch, ohne Jahr; 1979). Die Skalierung direkt auf das Deutsche anzuwenden wäre unangemessen. Briest weist beispielsweise auf unterschiedliche durchschnittliche Wortlängen hin: Die durchschnittliche Silbenlänge des englischen Wortschatzes betrage 1,42, die des deutschen aber 1,63 (Briest, 1974: 545). Der grundsätzliche Gedanke, dass Wort- und Satzlänge Hinweise auf die Textschwierigkeit darstellen, gilt jedoch auch für das Deutsche – wenn auch aus anderen Gründen als dies im Englischen der Fall ist. Dieser Gedanke veranlasste Mihm, die Skalierung von Flesch auf das Deutsche zu übertragen (Mihm, 1973: 120). Seine modifizierte Bewertung der Textverständlichkeit für deutschsprachige Texte geht aus Tabelle 31 (Seite 246) hervor.

Ähnliches gilt für den ebenfalls weit verbreiteten *Gunning Fog Index* (Gunning, 1952). Der Index berücksichtigt die Satzlänge und den Anteil von Wörtern mit mehr als zwei Silben, wobei bestimmte Wörter unberücksichtigt bleiben. Die Formel lautet:

Gunning Fog Index = (Wörter pro Satz + lange Wörter/Wörter*100) * 0,4

Ein niedriger Index weist auf verständliche und ein hoher auf unverständliche Texte hin. Wiederum für das Englische sind Zuordnungen zu Schul- bzw. Hochschuljahren getroffen worden für den Gunning Fog Index auf einer Skala von 6 bis 17. Eine Umrechnung auf das Deutsche ist mir nicht bekannt.

Tabelle 31: Beurteilung der Verständlichkeit deutscher Texte mit dem Flesch Index (nach Mihm)

Flesch-Index	Charakteristik	Typischer Text
-20 bis + 10	sehr schwer	wissenschaftliche Abhandlung
+10 bis 30	schwierig	Fachliteratur
30 bis 40	anspruchsvoll	Sachbuch, Roman (z. B. Buddenbrooks)
40 bis 50	normal	Roman (z. B. Stiller)
50 bis 60	einfach	Unterhaltungsliteratur (Karl May)
60 bis 70	leicht	Heftchenroman
70 bis 80	sehr leicht	Comics

Quelle: Mihm, 1973: 120

Tabelle 32: Lesetexte mit Fachbezug – Kennzahlen zur Textverständlichkeit

	Inflation	Geschwin-digkeit	Vergleich: Inflation	Vergleich: Radar
Wörter	489	453	457	405
Silben	1038	959	1058	924
Sätze	28	32	20	21
Silben pro Wort	2,12	2,11	2,31	2,28
Wörter pro Satz	17,46	14,15	22,85	19,28
Flesch-Index	**10**	**14**	**-12**	**-7**
lange Wörter	159	118	183	143
Gunning-Fog-Index	**20,0**	**16,1**	**25,2**	**21,8**

Anm: Flesch-Index: niedrige Zahl – schwieriger Text; Gunning-Fog-Index: hohe Zahl – schwieriger Text.

Tabelle 33: Lesetexte mit Fachbezug – Rangordnung nach Textverständlichkeit

Rang	Flesch-Index	Gunning-Fog-Index
1	Vergleichstext: Inflation	Vergleichstext: Inflation
2	Vergleichstext: Radar	Vergleichstext: Radar
3	Inflation	Inflation
4	Geschwindigkeit	Geschwindigkeit

Anm: vorderer Rangplatz → schwer verständlicher Text

In den Kernaussagen stimmten die Indices überein: Beide Indices weisen die Vergleichstexte als besonders schwer verständlich, und die Texte für die Leseverstehenstests als etwas weniger schwer verständlich aus. Nach Mihm sind die Vergleichstexte Inflation und Radar "sehr schwer" und typisch für eine wissenschaftliche Abhandlung, die Texte der Leseverstehenstests jedoch nur "schwierig" und typisch für Fachliteratur (siehe Tabelle 31 und

Tabelle 32). Auch der Gunning-Fog-Index weist die Vergleichstexte als schwieriger aus.

Meine eigene Einschätzung der Textverständlichkeit, die ich anhand des "Hamburger Verständlichkeitskonzepts" verdeutlichte, wird durch die quantitative Erfassung mit den beiden Indices in den Grundsätzen gestützt: Die Texte, welche in den Leseverstehenstests eingesetzt werden sollten, weisen einen ähnlichen Schwierigkeitsgrad auf, und sie sind verständlicher als Texte mit vergleichbarer inhaltlicher Ausrichtung aus dem Studienzusammenhang.

Die Items der Leseverstehenstests

Beim Sprechen oder Schreiben können schriftliche oder mündliche Äußerungen als Beleg für die Kompetenz bewertet werden. Beim Textverständnis ist man darauf angewiesen, dass die Kandidaten ihr Verständnis kommunizieren. Dabei können Störungen auftreten. Beispiele: Der Text wurde zwar verstanden, nicht aber die Frage oder die als richtig vorgesehene Antwort einer Multiple-Choice-Frage, die den Text paraphrasiert.

Zum Einfluss der Aufgabentypen auf das Testkonstrukt sind eine Reihe von Studien durchgeführt worden. In welcher Sprache sollten die Aufgabenstellungen verfasst sein? Shohamy (1984) untersuchte die unterschiedlichen Schwierigkeiten von Multiple-Choice Aufgaben in der Ausgangssprache und in der Fremdsprache. Sie stellte fest, dass die Kandidaten bessere Ergebnisse erzielten, wenn die Aufgabenstellungen in der Ausgangssprache verfasst waren. Shohamy bietet mehrere Erklärungsmodelle an: Es ist möglich, dass Items in der Fremdsprache auch unbekannte Wörter enthalten, welche eine zusätzliche Schwierigkeit darstellen. Die Verwendung der Ausgangssprache könnte auch zu einer Reduktion von Prüfungsangst und damit zu besseren Ergebnissen geführt haben. Zum anderen mutmaßt sie, dass die Kandidaten aus Items in der Ausgangssprache Hinweise für das Textverständnis erhalten haben. Für die Verwendung der Ausgangssprache spricht ihrer Meinung nach, dass dies eine authentische Sprachverwendungssituation darstellt: In der Regel würden sich Leser von fremdsprachlichen Texten Fragen in ihrer Ausgangssprache an den Text stellen.

Auf Sprachtests für den Hochschulzugang und Sprache im Studium trifft dies jedoch nicht zu. Typische Prüfungssituationen im Studium werden durchgehend in der Zielsprache durchgeführt. Aus organisatorischen Gründen wäre es schwierig, Items in der Ausgangssprache der Kandidaten anzubieten. Aus praktischen Erwägungen entschied ich mich dafür, die Aufgabenstellungen auf Deutsch zu formulieren, wenngleich die Muttersprache für die Erhebung des Textverständnisses möglicherweise geeigneter gewesen wäre.

Tabelle 34: Studie von Alderson und Urquhart – Einfluss unterschiedlicher Aufgabentypen

durchschnittliches Testergebnis	**Testversion 1 Lückentest**	**Testversion 2 Fragen zum Text**
höchstes Testergebnis	Text A	Text A
	Text B	Text D
mittleres Testergebnis	Text C	Text C
	Text D	**Text B**
niedrigstes Testergebnis	Text E	Text E

nach Alderson/Urquhart, 1985a; 1985b.

Unterschiedliche Aufgabentypen können durchaus zu unterschiedlichen Ergebnissen führen, auch wenn man von einem vergleichbaren Schwierigkeitsgrad der Texte ausgeht. Exemplarisch möchte ich auf die Untersuchung von Alderson und Urquhart verweisen, die die Auswirkungen von zwei Aufgabentypen verglichen: Lückentexte und Fragen zum Text. Alderson und Urquhart (1985b: 32-39) ließen gleiche (Fach-)Texte mit unterschiedlichen Aufgabentypen von Studenten bearbeiten, die in ihrem Heimatland bereits ein Studium absolviert hatten und die sich in Großbritannien in universitären Sprachkursen auf ein Postgraduiertenstudium vorbereiteten. Die erste Testversion enthielt fünf Lückentexte aus verschiedenen Disziplinen. Es wurden vollständige Wörter gestrichen; dabei war man nicht zufällig vorgegangen, sondern wählte die Wörter nach inhaltlichen Kriterien aus, so dass die einzelnen Items vor allem das Textverständnis prüfen sollten. Für die zweite Testversion wurde ein anderer Aufgabentyp gewählt: Die Kandidaten mussten kurze Fragen zum Text beantworten. Alderson und Urquhart beobachteten deutliche Auswirkungen des Aufgabentyps auf die Ergebnisse, obwohl zentrale Aussagen durch beide Aufgabentypen bestätigt werden konnten: Es wurde eine hohe Korrelation zwischen den Ergebnissen aus der ersten und aus der zweiten Version beobachtet ($r = 0{,}78$). Die Kandidaten erzielten in den Tests unterschiedliche Ergebnisse, die jedoch nicht nur von den Texten, sondern auch vom Aufgabentyp abhingen: Die Kandidaten erzielten bei einem Text in der ersten Testversion das zweitbeste Ergebnis, in der zweiten Testversion jedoch das zweitschlechteste und umgekehrt (siehe Tabelle 34, Seite 249). Diese Ergebnisse bestätigen die Annahme,

dass unterschiedliche Aufgabentypen auch zu unterschiedlichen Ergebnissen führen können. Es gibt eine Reihe weiterer Studien zu den Auswirkungen unterschiedlicher Aufgabentypen, die hier nicht vorgestellt werden (Alderson, 2000a; Bachman, 1990; Bachman/Lynch/Mason, 1995; Bachman/Palmer, 1982; Chapelle/Abraham, 1990; Douglas, 1998; Fulcher, 1996; Wigglesworth, 1997).

Aufgabentypen und Auswertungsmethoden spielen eine wichtige Rolle für die Schwierigkeit und die Konstruktvalidität der Items in Leseverstehenstests. Als allgemeine Empfehlung kann gelten, dass mehrere Aufgabentypen verwendet werden. Dies ist in den meisten standardisierten Sprachtests für den Hochschulzugang auch der Fall. Der TestDaF verwendet eine Zuordnungsaufgabe, dreigliedrige Multiple-Choice Aufgaben sowie trichotome Aufgaben (ja/nein/Text sagt dazu nichts). Beim IELTS ist das Spektrum noch breiter. Der Test enthält zusätzlich Aufgaben zum Informationstransfer (Schaubilder/Diagramme ausfüllen) und eine Zusammenfassung mit Lücken (Projektgruppe TestDaF, 2000; International English Language Testing System, 1999). Bei der DSH hängt die Anzahl der unterschiedlichen Aufgabentypen vom jeweiligen Ausrichter ab. Im DSH-Handbuch werden vielfältige Aufgabentypen angeregt, wobei nach geschlossenen, halboffenen und offenen Aufgabenstellungen unterschieden wird (FaDaF, 2001: 5/3).

Die Leseverstehenstests der Studien sollten einerseits den Grad des Textverständnisses der Kandidaten möglichst gut zum Ausdruck bringen und andererseits typische Aufgabentypen aus dem DSH-Leseverstehen enthalten. Wenn völlig andere Aufgabentypen eingesetzt würden, könnten andere Ergebnisse und Interpretationen der Ergebnisse die Folge sein. Empfehlungen mit Blick auf die DSH wären dann fragwürdig. Freilich ist das Spektrum der in DSH-Leseverstehenstests verwendeten Aufgabentypen wegen der fehlenden Standardisierung groß. Die DSH-Rahmenordnung ist in diesem Punkt sehr offen; im DSH-Handbuch werden Beispiele für mögliche Aufgabentypen gegeben, die jedoch nicht verbindlich sind. Folgende Aufgabentypen sind erfahrungsgemäß häufig anzutreffen: Multiple-Choice Fragen, dichotome Fragen (ja/nein, richtig/falsch), trichotome Fragen (ja/nein/nicht im Text), offene Fragen mit kurzen Antworten, Lückentexte, Zuordnungen und Aufgaben zum Informationstransfer z. B. zum Ausfüllen von Schaubildern.

Bei der Auswahl der Items für die Leseverstehenstests der Studien war weiter zu berücksichtigen, dass die Anzahl eher gering sein sollte. In nicht mehr als 90 Minuten sollten nicht nur Leistungen in verschiedenen Leseverstehenstests, sondern auch Vorkenntnisse und das Niveau der Deutschkenntnisse erhoben werden. Daher wurden Texte gewählt, deren Umfang *unter* der in der DSH üblichen Textlänge liegt. Auch die Anzahl der Items sollte eher gering sein.

Zu den Lesetexten "Inflation" und "Geschwindigkeit" entwarf ich jeweils zwei Tests mit unterschiedlichen Aufgabentypen, welche mit jeweils 17 Kandidaten erprobt wurden. Eine Version bestand weitgehend aus Multiple-Choice Fragen, eine andere aus Fragen mit kurzen Antworten und paraphrasierende Sätze mit Lücken. Zu diesen Aufgabentypen gibt es eine Vielzahl von Studien (Übersichten z. B. in Alderson, 2000a; Alderson/Clapham/Wall, 1995; Urquhart/Weir, 1998).

Ein Vorteil von Multiple-Choice Fragen liegt in der einfachen Auswertung. Problematisch ist, dass die Erstellung mehrerer Erprobungen bedarf und dass der Entscheidungsprozess, welcher zur Antwort führte, nicht nachvollziehbar ist. Da auch Zufallstreffer dabei sein können, hängt die Reliabilität von der Anzahl der Items ab. Es gibt jedoch viele Beispiele für reliable Tests mit Multiple-Choice Aufgaben (z. B. TOEFL).

Fragen mit kurzen Antworten und paraphrasierende Sätze mit Lücken erlauben eher Rückschlüsse über den Verständnisprozess. Im Gegensatz zu Multiple-Choice Aufgaben müssen die Kandidaten bei Fragen mit kurzen Antworten selbst eine Antwort formulieren. Probleme bei der Bewertung der Antwort entstehen, wenn unvorhergesehene Antworten gegeben werden oder wenn größere Passagen mit teilweise passenden Antworthinweisen aus dem Text abgeschrieben werden. Daher sind bei Fragen mit kurzen Antworten ebenfalls Erprobungen nötig. Wenn die Antwort nicht direkt aus dem Text entnommen werden kann, wird das Item schwieriger. Denkbar ist auch, dass zum Konstrukt von derartigen Items nicht nur Leseverstehen, sondern auch Schreiben zu zählen ist.

Ich entschied mich für die Version, welche Fragen mit kurzen Antworten und paraphrasierende Sätze mit Lücken enthält. Grund war auch, dass die Tests keine Auswirkungen für die Kandidaten hatten. Die Teilnahme war eine Gefälligkeit, oder sie wurde von der Lehrkraft einfach angeordnet. Vor diesem Hintergrund schienen Aufgabentypen, bei denen die Kandidaten selbst eine Antwort formulieren mussten, zu einer intensiveren Auseinandersetzung mit dem Text anzuregen als Multiple-Choice Fragen, die möglicherweise zufällig abgehakt werden. Ein weiterer Vorteil der Fragen mit kurzen Antworten lag in der Möglichkeit, eine mehrstufige Punktevergabe anzuwenden.

Beantworten Sie die folgenden Fragen mit Stichworten.

1) Was versteht man unter einer Inflation?
   .....
2) Was wird mit der Zahl 4,3 % genau ausgedrückt? (Zeile 27)
   .....
3) Welche Auswirkungen hat eine Inflation normalerweise auf die Kaufkraft?
   .....
4) Welche Auswirkungen hat eine Deflation normalerweise auf die Lohnentwicklung?
   .....
5) Wie verhält sich die umlaufende Geldmenge bei einer Deflation?
   .....
6) Wann könnte der Staat ein Interesse an einer Deflation haben?
   .....
7) Welche Rolle spielt die Deflation derzeit in Europa?
   .....
8) Worauf bezieht sich "sämtlicher"? (Zeile 29)
   .....
9) Worauf bezieht sich "das"? (Zeile 52)
   .....
10) Worauf bezieht sich das erste und worauf das zweite "die"? (Zeile 79)
    1. "die"..... 2. "die" .....

Füllen Sie die Lücken aus.

a) Der Warenkorb wird einmal mit den Preisen des ....., zum anderen mit den Preisen des Berichtsjahres bewertet.
b) Laut Lexikontext war die Inflationsrate in den USA in den neunziger Jahren ..... als in Deutschland.
c) Wenn die Preissteigerung zurückgeht, spricht man von .....
d) Die Vorlieben der Käufer und die Qualität der Produkte ändern sich. Daher muss auch ..... geändert werden.
e) Wenn die Preise im Warenkorb über einen längeren Zeitraum zugenommen haben, liegt ..... vor.

Abbildung 25: Leseverstehenstext "Inflation" – Items

Der Schwierigkeitsgrad der Items sollte sich unterscheiden. Konnten einige Fragen durch Textexzerpte beantwortet werden, so waren andere nur durch Kombination von mehreren Textstellen zu beantworten. Weil nur wenige Items eingesetzt wurden, sollten die Items jedoch nicht zu einfach und nicht zu schwierig sein. Dafür berechnete ich jeweils einen Schwierigkeitsindex und überarbeitete oder strich einzelne Items (Brown, 1996: 49-92; Lienert/Raatz, 1994: 73-78). Die in der Studie eingesetzten Items zu den Texten "Inflation" und "Geschwindigkeit" gehen aus der Abbildung 25 und der Abbildung 26 (Seite 252 bzw. 254) hervor.

Es wurden folgende Aufgabentypen verwendet: Offene Fragen zum Text, die mit Stichwörtern oder mit einem Satz beantwortet werden können, sowie Sätze mit Lücken, in denen der Inhalt des Textes paraphrasiert oder kommentiert wird. In allen Leseverstehenstests mit geringem Fachlichkeitsgrad wurden auch Textbezüge erfragt. Art und Anzahl der Items gehen aus Tabelle 35 (Seite 254)hervor. Die Tests "Inflation" und "Geschwindigkeit" enthalten jeweils eine vergleichbare maximale Punktzahl. Wie bereits bei den Grammatiktests, wurden auch im Fall der Leseverstehenstests angeglichene Testformen verglichen, d. h. die Ergebnisse wurden linear in eine gemeinsame Skala (Prozentwerte) transformiert.

Bewertet wurde der Inhalt, nicht die Sprache. Ich gehe davon aus, dass die Ergebnisse in den vorgestellten Leseverstehenstests Interpretationen mit Blick auf das Leseverstehen von Fachtexten mit erkennbarem, aber niedrigem Fachlichkeitsgrad ermöglichen. Zunächst geht es jedoch um die Vergleichbarkeit von *Tests* zum Leseverstehen. Ob die in den Tests demonstrierten Fähigkeiten ein Abbild des tatsächlichen Verständnisgrads darstellen, wird nicht explizit erhoben, da es für die Zielsetzung der Studie nur von untergeordneter Bedeutung ist.

Beantworten Sie die folgenden Fragen in Stichworten.

1) Welche Messgrößen werden benötigt, um die Geschwindigkeit eines Fahrzeugs zu bestimmen?
   .....
2) Wie erhält man den zurückgelegten Weg bei der Geschwindigkeitsmessung mit Induktionsschleifen?
   .....
3) Was macht die Polizei, um Autofahrer zu identifizieren, die zu schnell fahren?
   .....
4) Wie kann man näherungsweise eine Momentangeschwindigkeit mit Induktionsschleifen messen?
   .....
5) Was ist die Aufgabe des "Empfängers" einer Laserpistole? (Zeile 51)
   .....
6) Wie wird die Strecke bei der Geschwindigkeitsmessung mittels Laser bestimmt?
   .....
7) Wie viele Messungen sind zur Bestimmung der Geschwindigkeit mittels Laser notwendig?
   .....
8) Wann ist bei einer Radarmessung eine Veränderung der Frequenz festzustellen?
   .....
9) Wessen Frequenz ändert sich? (Zeile 77)
   .....
10) Worauf bezieht sich "dazu"? (Zeile 21)
    .....
11) Worauf bezieht sich "sie"? (Zeile 26)
    .....

Füllen Sie die Lücken aus.

a) Die mittlere Geschwindigkeit während der gesamten Fahrt bezeichnet man auch als
   .....
b) Die Erzeugung einer elektrischen Spannung mit Hilfe veränderlicher magnetischer Felder nennt man
   .....
c) Verändert sich ..... der elektro-magnetischen Wellen, kann die Geschwindigkeit eines Körpers bestimmt werden.
d) Zur Messung der Geschwindigkeit eines fahrenden Autos mit Induktionsschleifen muss man die Entfernung der beiden Schleifen kennen und außerdem
   .....

Abbildung 26: Leseverstehen "Geschwindigkeit" – Items

Tabelle 35: Leseverstehenstests – Items

"Inflation"
10 Fragen mit kurzen Antworten 5 Paraphrasierungen des Inhalts mit Lücken
"Geschwindigkeitsmessung"
11 Fragen mit kurzen Antworten 4 Paraphrasierungen des Inhalts mit Lücken

6.1.3. Erhebung der Deutschkenntnisse

In Studien zur Rolle von Vorkenntnissen auf die Ergebnisse in Leseverstehenstests wurden die Sprachkenntnisse mit unterschiedlichen Methoden erhoben. Clapham griff beispielsweise auf Grammatiktests zurück, welche damals Teil des IELTS waren (Clapham, 1996). Wie in Kapitel 4 gezeigt wurde, eignen sich Grammatiktests zwar zur Differenzierung zwischen den produktiven Grammatikkompetenzen der Teilnehmer. Ein Abbild verschiedener sprachlicher Fertigkeiten bieten sie jedoch nicht notwendigerweise.

In der vorliegenden Studie wurde das Niveau der Deutschkenntnisse anhand zweier C-Tests mit 40 Items erfasst. C-Tests wurden eingesetzt, weil sie sich in mehreren Studien als geeignetes Testinstrument zur Erhebung "allgemeiner Sprachkompetenz" erwiesen. Wenn man nach Bachman (1990) bei dem Konstrukt Sprachkompetenz zwischen "*organizational*" und "*pragmatic competence*" unterscheidet, sind Leistungen in C-Tests eher als Indikator für "*organizational competence*" anzusehen (siehe Kapitel 2.1, "Modelle von Sprachkompetenz", Seite 35 ff). Dies wurde in einer Reihe von Studien bestätigt (Zusammenfassungen in: Coleman/Grotjahn/Raatz, 2002; Grotjahn 1992; 1994; 1995; 1996; 2002). C-Tests bestehen aus zufällig ausgewählten Texten und die Lücken stellen eine repräsentative Auswahl der Wörter aus dem Text dar. In Studien sind hohe Korrelationen von C-Tests mit Tests anderer Sprachfertigkeiten aufgetreten. Hohe Korrelationen wurden auch mit externen Kriterien beobachtet (z. B. Selbsteinstufung der Lerner, Bewertung durch Lehrkräfte). Der wahrgenommene Schwierigkeitsgrad von C-Tests nimmt bei steigender Fremdsprachenkompetenz ab (und umgekehrt). Obwohl die meisten Sprachverarbeitungsstrategien, die von C-Tests hervorgerufen werden, der Mikroebene zuzuordnen sind, wurde in psycholinguistischen Studien, dass (vor allem bei fortgeschrittenen Sprechern) von C-Tests auch höhere Sprachverarbeitungsstrategien hervorgerufen werden.

Den Kandidaten wurde vor der Bearbeitung der Leseverstehenstests ein Fragebogen zu den Vorkenntnissen (Abbildung 28, Seite 258) und zwei C-Tests vorgelegt (Abbildung 27, Seite 256). Die C-Tests wurden mit kleinen Abweichungen nach der klassischen C-Test-Methode konzipiert: In kurzen Texten wird die letzte Hälfte von jedem zweiten

Wort gestrichen. Idealerweise hätten die Kandidaten mehr als nur zwei C-Tests mit je 20 Lücken ausgefüllt, aber zusammen mit den Leseverstehenstests und dem Fragebogen wäre das Testpaket zu umfangreich geworden. Die verwendeten C-Tests waren bereits in Aufnahmeprüfungen zum Studienkolleg eingesetzt worden und hatten sich dort bewährt. Insgesamt waren dort vier C-Tests eingesetzt worden. Die ausgewählten Tests waren besonders geeignet, weil sie deutlich zwischen den Leistungen der Kandidaten differenzierten und einen mittleren Schwierigkeitsindex aufwiesen (Brown, 1996: 64-86). Im Folgenden betrachte ich die beiden C-Tests mit insgesamt 40 Items als einen Test und verwende daher den Singular: der C-Test.

Füllen Sie die Lücken aus!

Bedeutung der Lesefähigkeit
Die Lesefähigkeit trägt ihren Wert natürlich in sich, hat aber auch ökonomische Auswirkungen. Erwac_____ Leser, d_____ besser le_____ als d_____ Durchschnitt, üb_____ mit größ_____ Wahrscheinlichkeit gutbe_____ Berufe a_____. Die wach_____ Spezialisierung i_____ der Gesell_____ erfordert me_____ Bildung, ei_____ Forderung, d_____ vor al_____ an d_____ Schulen geri_____ wird. Du_____ die erhö_____ Anforderungen a_____ das Bildungsniveau, die heute in den westlichen Gesellschaften gestellt werden, ist die Lesefähigkeit des Einzelnen immer wichtiger geworden.

Naturkatastrophen
Neben den plötzlich auftretenden Naturkatastrophen gibt es natürliche Risiken, die kontinuierlich vorhanden und schwer erkennbar sind: etwa die natürlich vorkommende Radioaktivität oder natürliche toxische Metallvorkommen in der Umwelt. Zu_____ können ein_____ natürliche Ris_____ durch d_____ Eingriffe d_____ Menschen verschlimmert wer_____: etwa Überschw_____ aufgrund d_____ Zerstörung v_____ Wäldern. F_____ die Erfor_____ dieser Gefa_____ sind des_____ die grundl_____ Erkenntnisse d_____ Umweltwissenschaften v_____ zentraler Bede_____. Die schwe_____ Risiken du_____ Naturkatastrophen best_____ in den wirtschaftlich noch wenig entwickelten Staaten. Dies liegt teils an den klimatischen Bedingungen der Tropen, teils an der Lage innerhalb geologischer Schwäche- oder Gefahrenzonen und schließlich an der noch gering ausgebauten Infrastruktur bezüglich Schutzmaßnahmen für Mensch und Umwelt.

Abbildung 27: C-Test zur Erhebung der Deutschkenntnisse (vollständige Erhebung im Anhang)

6.1.4. Erhebung der Vorkenntnisse

Mehrere Studien zur Rolle der Fachkenntnisse, über die in Kapitel 5.2 (Seite 199 ff) bereits berichtet wurde, litten unter der ungenauen Erhebung der Vorkenntnisse. Häufig wurde im Resümee darauf hingewiesen, dass der Erhebung der Vorkenntnisse in Folgestudien mehr Beachtung zukommen sollte. Daher legte ich auf die Ermittlung der Vorkenntnisse einen besonderen Wert. Wenn man mit Bernhardt (1991b) Vorkenntnisse nach der Art des Erwerbs unterscheidet, geht es hier um Fachkenntnisse, die vor allem durch formelle Bildung erworben wurden. Individuelles oder kulturelles Wissen wurde nicht ausdrücklich erhoben. Die Vorkenntnisse wurden anhand folgender Kriterien ermittelt:

- Erhebung der Vorkenntnisse durch Fragen nach Schlüsselbegriffen vor dem Lesen des Textes (Variable: BEGRIFFE),
- Einschätzung der Vorkenntnisse durch die Kandidaten nach dem Lesen des Textes (Variable: BEKANNT).
- Kurszuordnung im Studienkolleg bzw. Studienziel (Variable: KURS),

Den Kandidaten wurde vor der Bearbeitung der Leseverstehenstests "Inflation" und "Geschwindigkeitsmessung" ein Kurzfragebogen zur Erfassung der Vorkenntnisse vorgelegt. In diesem Kurzfragebogen wurden außerdem einige biografische Hintergrundinformationen erfragt, welche möglicherweise einen Einfluss auf die Vorkenntnisse haben: Die Kurszuweisung am Studienkolleg, das angestrebte Studienfach, Informationen über Studienerfahrungen sowie Informationen über eine etwaige Berufstätigkeit. Nach der Bearbeitung der Leseverstehenstests sollten die Kandidaten noch einmal Auskunft über ihre Vorkenntnisse zum Thema geben.

Vorkenntnisse nach Kenntnis der Schlüsselbegriffe vor dem Lesen (BEGRIFFE)

Die Variable BEGRIFFE wurde vor dem Lesen der Texte erhoben. Die Testteilnehmer wurden gebeten, Fragen zu jeweils zwei Schlüsselbegriffen der Texte zu beantworten: Sie sollten die Begriffe "Inflation", "Deflation", "Laser" und "Radar" erklären.

Wie aus Tabelle 36 hervorgeht, waren deutlich mehr Kandidaten mit einem der Begriffe "Inflation" oder "Deflation" vertraut als mit einem der Begriffe "Radar" oder "Laser". Zur Codierung: Wenn Kandidaten einen der Begriffe erklären konnten, wurde die Variable BEGRIFFE mit 1 codiert, sonst mit 0. Die Variable wurde also zweimal erhoben: Für den Text "Inflation" und für den Text "Geschwindigkeit".

Was ist ein "**Laser**"?

.....

Was versteht man unter "**Radar**"?

.....

Was spricht man von "**Inflation**"?

.....

Wann liegt eine "**Deflation**" vor?

.....

Abbildung 28: Erhebung der Vorkenntnisse nach Kenntnis der Schlüsselbegriffe vor dem Lesen (vollständige Erhebung im Anhang)

Tabelle 36: Vorkenntnisse nach Kenntnis der Schlüsselbegriffe vor dem Lesen (Variable BEGRIFFE) – Häufigkeiten

	Ja	**nein**	**Gesamt**
Vorkenntnisse zum Text "Inflation" (Kenntnis der Schlüsselbegriffe "Inflation" und/oder "Deflation")	313	206	519
Vorkenntnisse zum Text "Geschwindigkeit" (Kenntnis der Schlüsselbegriffe "Laser" und/oder "Radar"	205	294	499

Vorkenntnisse laut Selbstauskunft nach dem Lesen (BEKANNT)

Nach dem Lesen und dem Bearbeiten der Items sollten die Kandidaten erklären, ob ihnen der Inhalt des Textes bereits bekannt war oder nicht (Abbildung 29).

Die Verteilung der Antworten geht aus der Tabelle 37 hervor. Es gab wiederum mehr Kandidaten, denen der Text "Inflation" bereits vertraut war, und beim Lesetext "Geschwindigkeit" mehr Kandidaten, für die der Inhalt neu war. Allerdings ist das Verhältnis im Vergleich zur Erhebung vor dem Lesen anders: Bei der Erhebung der Vorkenntnisse nach dem Lesen gaben jeweils mehr Kandidaten an, dass ihnen das Thema des Textes vertraut war, als durch die Befragung vor dem Lesen deutlich wurde.

Einerseits dürfte die Einschätzung der Vorkenntnisse laut Selbstauskunft nach dem Lesen die verlässlichste Erhebungsmethode sein. Schließlich haben die Kandidaten den Text gelesen und können sich nun dazu äußern, ob ihnen der Inhalt bereits vertraut war oder nicht. Andererseits gibt es auch bei dieser Erhebung Störgrößen: Möglicherweise halten die Kandidaten die eine oder andere Antwort für geschickter, so dass die Antwort unehrlich sein könnte. Möglicherweise spielt das Niveau der Deutschkenntnisse eine wichtige Rolle für das Textverständnis und damit für den subjektiven Eindruck der Vertrautheit mit dem Thema.

Einschätzung der Texte:

Welche Aussage trifft auf Sie zu?

Die Informationen aus dem Text "Inflation" sind mir ...

❒ teilweise schon bekannt

❒ eher unbekannt

Die Informationen aus dem Text "Geschwindigkeit" sind mir ...

❒ teilweise schon bekannt

❒ eher unbekannt

Abbildung 29: Erhebung der Vorkenntnisse laut Selbstauskunft nach dem Lesen

Tabelle 37: Vorkenntnisse laut Selbstauskunft nach dem Lesen (Variable BEKANNT) – Häufigkeiten

	ja	**Nein**	**gesamt**
Informationen aus dem Text "Inflation" bereits bekannt?	317	200	517
Informationen aus dem Text "Geschwindigkeit" bereits bekannt?	236	260	496

Vorkenntnisse nach Kurszuordnung im Studienkolleg bzw. Studienziel (KURS)

Die ausländischen Studienbewerber werden den Kursen im Studienkolleg nach dem gewünschten Studienziel zugewiesen. Wer ein technisches Studium anstrebt, wird dem Technikkurs (T-Kurs) zugeordnet, wer ein wirtschaftliches Studium anstrebt, wird dem Wirtschaftskurs (W-Kurs) zugeordnet, wer Medizin oder verwandte Studiengänge belegen möchte, wird dem Medizinkurs (M-Kurs) zugeordnet, als Vorbereitung auf geisteswissenschaftliche Studiengänge gibt es den G-Kurs usw. In den Technikkurs werden auch ausländische Studienbewerber aufgenommen, die Informatik oder Architektur studieren möchten. Im Wirtschaftskurs befinden sich auch angehende Design- oder Kommunikationsdesign-Studierende. Das Spektrum der angestrebten Studiengänge ist also breiter als dies die Unterscheidung zwischen dem Technik- und dem Wirtschaftskurs auf den ersten Blick ahnen lässt.

Um den Einfluss der Vorkenntnisse anhand der Kurszugehörigkeit bestimmen zu können, sind lediglich Kandidaten aus T-Kursen bzw. aus W-Kursen von Interesse. Diese stellen auch die größten Gruppen (Tabelle 38). Die Teilnehmer aus den anderen Kursen spielen bei der Bestimmung der Vorkenntnisse nach Studienziel bzw. nach Kurszuordnung keine Rolle, da nur wirtschaftliche und technische Themen gewählt wurden. Die Ergebnisse dieser Kandidaten werden nicht berücksichtigt.

Die Validität der Variablen KURS mit Blick auf die Vorkenntnisse zum Thema der Texte ist eingeschränkt, denn die Variable KURS bezieht sich nicht auf die Vertrautheit mit dem konkreten Thema des Texts, der im Leseverstehenstest eingesetzt wurde. Es ging allenfalls um eine allgemeine Affinität zum Fachbereich, die sich in dem Studienziel bzw. der Kurszuordnung manifestieren könnte. Der Studienwunsch könnte das Ergebnis von einschlägigen Vorerfahrungen bzw. einer früheren Weichenstellung sein (z. B. dem Besuch des technischen Zweigs auf der Schule). Im Studienkolleg sind die Kandidaten bereits mit einigen Fachinhalten vertraut gemacht worden, auch dies könnte dazu beitragen, dass Vorkenntnisse vorhanden sind.

Tabelle 38: Vorkenntnisse nach Kurszuordnung im Studienkolleg bzw. Studienziel (Variable KURS) – Häufigkeiten

Kurszugehörigkeit/ Studienziel	LV-Inflation n	LV-Geschw. n
T-Kurs (Technische Studiengänge)	193	192
W-Kurs (Wirtschaftliche Studiengänge)	159	141
G-Kurs (Geisteswissenschaftl. Studiengänge)	109	108
M-Kurs (Medizin)	55	55
Gesamt	516	495

Tabelle 39: Vorkenntnisse – Übereinstimmungskoeffizienten und Korrelationskoeffizienten zwischen den drei Variablen

	LV "Inflation" n = 350		LV "Geschwindigkeit" n = 331	
	Übereinstimmungskoeffizient	Korrelationskoeffizient (Spearman)	Übereinstimmungskoeffizient	Korrelationskoeffizient (Spearman)
KURS und BEGRIFFE	72,9 %	,499; $p < 0,01$	50,2 %	,024; n. sig.
BEGRIFFE und BEKANNT	80,3 %	,573; $p < 0,01$	65,9 %	,330; $p < 0,01$
BEKANNT und KURS	64,0 %	,320; $p < 0,01$	66,8 %	,330; $p < 0,01$

Die Einschränkungen werden durch die Befunde anderer Studien unterstützt. Es zeigte sich, dass das zukünftige Studienfach allein kein verlässlicher Hinweis auf Vorkenntnisse ist (z. B. Alderson/Urquhart 1985a; 1985b). Denkbar ist einerseits, dass Kandidaten, die ein technisches/naturwissenschaftliches Studium anstreben, mit dem Thema Geschwindigkeitsmessung nicht vertraut sind. Denkbar ist auch, dass Kandidaten mit einem anderen Studienziel bereits Vorkenntnisse besitzen. Die Variable wurde dennoch in die Analyse einbezogen, weil Sprachtests mit Fachbezug für Kandidaten nach zukünftigem Studienfach erstellt werden. Die Information, welche Rolle die Variable Studienfach spielen, ist für die Studie relevant, auch wenn KURS nicht vorbehaltlos als Indikator für Vorkenntnisse interpretiert werden kann.

Wie reliabel sind die drei Variablen? Dazu wurden Übereinstimmungskoeffizienten und Korrelationskoeffizienten berechnet (siehe Tabelle 39). Die Übereinstimmungskoeffizienten zeigen den prozentualen Anteil der Kandidaten, die der gleichen Gruppe zugewiesen werden (Brown, 2001: 171). Korrelationskoeffizienten verdeutlichen den Zusammenhang zwischen zwei Variablen. Besonders gering ist die Übereinstimmung zwischen den Variablen KURS und BEGRIFFE mit Bezug auf GESCHWINDIGKEIT. Das bedeutet, dass es keinen Zusammenhang gibt zwischen dem Wunsch, ein technisches Studium aufzunehmen, und der Kenntnis der Begriffe "Radar" oder "Laser". Die Übereinstimmungen zwischen den übrigen Variablen sind höher. Wenn man davon ausgeht, dass die drei Variablen zu den Vorkenntnissen eigentlich das gleiche Konstrukt erfassen sollten, ist die Übereinstimmung zwischen den Variablen als gering anzusehen. Dies unterstützt die Vorgehensweise in der vorliegenden Studie: Vorkenntnisse lassen sich kaum mit einer einzigen Variablen erfassen.

Aus Korrelationen der Variablen untereinander und aus Korrelationen der Variablen zu den Vorkenntnissen mit C-TEST lassen sich weitere Hinweise gewinnen. Die Korrelationstabellen sind im Zusammenhang mit den Regressionsanalysen dargestellt (Tabelle 50 und Tabelle 52, Seite 284 und 287). Der niedrigste Zusammenhang ist jeweils zwischen C-TEST und BEKANNT zu verzeichnen ($r_s = 0{,}136$, $p < 0{,}5$ bei INFLATION und $r_s = -0{,}090$, nicht signifikant bei GESCHWINDIGKEIT). Das bedeutet, dass die Variable BEKANNT Deutschkenntnisse (gemessen mit C-TEST) nicht oder nur in sehr geringem Maße abbildet. Das Verhältnis der übrigen Variablen zu Deutschkenntnissen ist weniger deutlich. Dieser Aspekt wird in Kapitel 6.2.1 (Seite 268 ff) wieder aufgegriffen, wenn es um den Einfluss der Vorkenntnisse auf die Ergebnisse in den Leseverstehenstests mit Fachbezug geht.

Man hätte erwarten können, dass Studienbewerber, die technische Studiengänge belegen möchten, mit den Begriffen "Radar" und "Laser" vertrauter sind als Studienbewerber, die ein Studium der Wirtschaftswissenschaften anstreben. Möglicherweise erfasst die Variable BEGRIFFE auch Sprachkenntnisse und erst in zweiter Linie Vorkenntnisse. Die Variable BEGRIFFE korreliert dementsprechend höher mit den Leistungen im C-Test als die übrigen Variablen zu den Vorkenntnissen.

6.1.5. Hypothesen

Die Fragestellungen der Studie lauten:

Fragestellung 1
Erzielen ausländische Studienbewerber bessere Ergebnisse in Leseverstehenstests mit Fachbezug, wenn sie über Vorkenntnisse verfügen?

Fragestellung 2
Spielen die Fremdsprachkenntnisse oder die etwaigen Vorkenntnisse zum Thema die größere Rolle für die Ergebnisse in fremdsprachlichen Leseverstehenstests mit Fachbezug?

Fragestellung 3
Hängt der Einfluss der Vorkenntnisse vom Niveau der Deutschkenntnisse ab? Lassen sich sprachliche Schwellen ermitteln und beschreiben, bei denen sich der Einfluss der Vorkenntnisse zum Thema verändert?

Wenn die Untersuchungsergebnisse von Studien zu Sprachtests mit Fachbezug (Kapitel 5.2, Seite 207 ff) auch auf die vorliegende Studie übertragbar sind, dürften Vorkenntnisse dazu führen, dass ausländische Studienbewerber bessere Ergebnisse in Leseverstehenstests mit Fachbezug führen. Ebenso ist zu erwarten, dass das Niveau der Fremdsprachenkompetenz einen größeren Einfluss für die Ergebnisse in fremdsprachlichen Leseverstehenstests mit Fachbezug spielen als das Vorhandensein von Vorkenntnissen. Für offen halte ich die Frage nach der "Doppelten Schwellenhypothese": Die existierenden Studien zu diesem Thema leiden unter methodischen Schwächen, die Ergebnisse dieser Studien müssen mit Vorbehalten betrachtet werden (siehe Kapitel 5.2, Seite 220 ff). Denkbar ist, dass tatsächlich sprachliche Schwellen erkennbar sind, bei denen sich der Einfluss der Vorkenntnisse ändert. In diesem Fall müsste die Erwartung der "Doppelten Schwellenhypothese" bei der Testerstellung berücksichtigt werden (Überlegungen dazu in Kapitel 5.2, Seite 225 ff). Doch auch für den Fall, dass sich die Annahmen der "Doppelten Schwellenhypothese" nicht manifestieren sollten, sind die

Erkenntnisse für die Testerstellung und die Interpretation der Testergebnisse von Interesse.

6.2. Ergebnisse und Diskussion

Übersicht: Kapitel 6.2

In diesem Kapitel werden die Ergebnisse der Studie vorgestellt und eingeordnet. Am Anfang stehen die statistischen Kennwerte und Korrelationen der beiden Leseverstehenstests mit Fachbezug. Es folgen Ergebnisse zur Rolle der Vorkenntnisse (Kapitel 6.2.1)und zum Einfluss der Vorkenntnisse im Vergleich zu den Deutschkenntnissen (Kapitel 6.2.2). Schließlich wird untersucht, ob sich sprachliche Schwellen ausmachen lassen, bei denen sich der Einfluss der Vorkenntnisse ändert (Kapitel 6.2.3).

Leseverstehenstests mit Fachbezug im Vergleich

Ergebnisse: An den Tests "Inflation" und "Geschwindigkeit" nahmen 486 Kandidaten teil. Statistische Kennwerte sind in Tabelle 40 dargestellt. Im Mittel erzielten die Kandidaten im Leseverstehenstest "Geschwindigkeit" ein etwas höheres Ergebnis als im Leseverstehenstest "Inflation". Der Unterschied zwischen den mittleren Ergebnissen ist nicht signifikant, wie mit einem *t*-Test für abhängige Stichproben mit den Ergebnissen der Kandidaten gezeigt werden kann, die an beiden Tests teilnahmen.

Die Streuung der Ergebnisse ist beim Leseverstehenstest "Geschwindigkeit" etwas breiter als bei "Inflation" ("Geschwindigkeit": s = 22,9; "Inflation": s = 20,8). Der die Gesamtzahl in zwei Hälften teilende Wert liegt etwas über dem arithmetischen Mittel, was darauf hindeutet, dass die Ergebnisse im Leseverstehenstest "Geschwindigkeit" im oberen Bereich etwas dichter beieinander liegen und im unteren Bereich weiter auseinander liegen (siehe Tabelle 40, Seite 266 und Abbildung 30, Seite 267).

Tabelle 40: Leseverstehenstests mit geringem Fachlichkeitsgrad – statistische Kennwerte

	LV-Inflation	**LV-Geschwindigkeit**
Anzahl der Kandidaten; N (Teilnehmer an beiden Tests; n)	516 (486)	495 (486)
Anzahl der Items	15	15
Mittelwert der Ergebnisse (*AM*) in % (*AM* der Teilnehmer an beiden Tests)	52,4 (51,5)	52,7 (52,8)
Vergleich der Mittelwerte (*t*-Test für abhängige Stichproben)	$t(498) = 1{,}490$; nicht signifikant	
Standardabweichung (Streuungsmaß; s) des Prozentwerts	20,5	22,6
Median (*Md*) als %	52,9	53,0
Maximum – Minimum als %	4-100	0-100
Korrelationskoeffizient nach Pearson (r); Signifikanzniveau	$r = 0{,}584$; $p < 0{,}01$	

Zusammenfassung und Diskussion: Der Vergleich der Ergebnisse in den Leseverstehenstests "Inflation" und "Geschwindigkeit" lässt folgende Interpretationen zu. Es gibt einen hoch signifikanten, mittleren Zusammenhang zwischen den Ergebnissen in den beiden Leseverstehenstests. Der gemeinsame Varianzbereich beträgt 29 Prozent ($r^2 = 0{,}2937$).

Warum ist der Zusammenhang nicht höher? Beabsichtigt war, zwei Leseverstehenstests zu konzipieren, die zwar über einen unterschiedlichen Fachbezug verfügen, die aber mit Blick auf den Schwierigkeitsgrad der Texte und der Items durchaus als Paralleltests gelten könnten. Von Paralleltests könnte man aber nur sprechen, wenn die Korrelation höher gewesen wäre. Die beobachtete Korrelation ist zwar ausgeprägt, sie liegt jedoch nicht in einem hohen, sondern eher in einem mittleren Bereich. Zur Erklärung dieses Phänomens gibt es mehrere mögliche Ansätze: Erstens könnte – wie in den Hypothesen formuliert – der unterschiedliche Fachbezug dazu führen, dass der Zusammenhang zwischen en Ergebnissen in beiden Tests nicht so ausgeprägt ist. Dieser Frage gehe ich in Kapitel 6.2.1 nach. Zweitens könnte es sein, dass es ohnehin schwierig ist, zwei Leseverstehenstests mit gleichem Schwierigkeitsgrad zu konstruieren. Es ist möglich, dass

die Leistungen im Leseverstehen variieren, auch wenn Texte und Items mit vergleichbarem Schwierigkeitsgrad vorgelegt werden. Möglicherweise variieren sie so stark, dass höhere Korrelationen bei einer großen Zahl von Kandidaten nicht zu erwarten sind (vgl. Alderson, 1984; 1999; 2000: 32-85; Farr/Carey/Tone, 1985; Grabe, 1991; Rost, 1993; Urquhart/Weir, 1998; Weir/Huizhong/Yan, 2000).

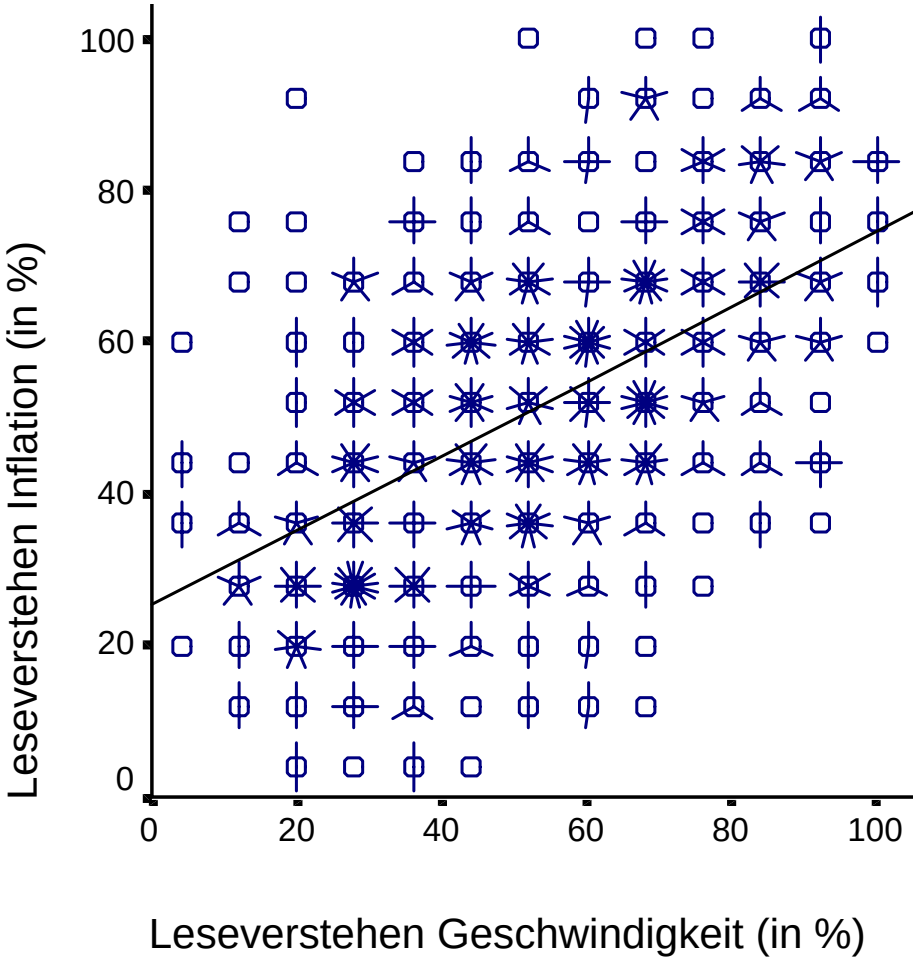


Abbildung 30: Ergebnisse in Leseverstehenstests "Inflation" und "Geschwindigkeit" (Sonnenblumen-Streudiagramm mit linearer Regressionsgeraden; n = 499)

6.2.1. Vorkenntnisse und Leseverstehenstests mit Fachbezug

Die Ergebnisse zum Einfluss der Vorkenntnisse werden in drei Gruppen dargestellt: Zunächst geht es um den Einfluss nach der Variablen Kurszugehörigkeit/Studienziel (KURS) auf die Leistungen in den Leseverstehenstests "Inflation" und "Geschwindigkeitsmessung". Anschließend werden die Leistungen in den Leseverstehenstests mit den Vorkenntnissen der Kandidaten verglichen, die aufgrund der Fragen zu den Schlüsselbegriffen der Texte erhoben wurden (BEGRIFFE). Schließlich werden die Aussagen zu Vorkenntnissen, welche die Kandidaten nach dem Lesen äußerten, mit den Ergebnissen in den Leseverstehenstests in Relation gesetzt (BEKANNT).

Einige Hinweise zur Statistik: Die Signifikanztests in diesem Kapitel sind Ergebnisse einfaktorieller Varianzanalysen. Der Levene-Test auf Gleichheit der Fehlervarianzen wies in jedem Fall eine Varianzenhomogenität aus. Die Normalverteilung, welche graphisch und rechnerisch annähernd bestimmt wurde, war jedoch nicht immer eindeutig. Daher wird bei den folgenden Varianzanalysen ein höheres Signifikanzniveau verlangt.

Vorkenntnisse nach KURS

Ergebnisse: Die Testteilnehmer aus W-Kursen erzielten im Leseverstehenstest "Inflation" durchschnittlich 61 Prozent. Das Ergebnis der Kandidaten aus T-Kursen war mit 48 Prozent deutlich niedriger. Zum Vergleich der Mittelwerte wurde eine einfaktorielle Varianzanalyse durchgeführt. Der Unterschied zwischen den Mittelwerten ist hoch signifikant (Tabelle 41, Seite 271).

Auch beim Leseverstehenstest "Geschwindigkeit" erzielten die Kandidaten, deren Studienziel eine Nähe zum Thema des Textes aufweist, im Mittel ein höheres Ergebnis. Die Kandidaten aus den T-Kursen erzielten 59 Prozent, diejenigen aus dem W-Kurs nur 52 Prozent. Der Unterschied zwischen den Ergebnissen ist ebenfalls hoch signifikant.

Aus dem "Mittelwert der Diagonalen" lässt sich ein weiterer Hinweis darauf gewinnen, dass Kandidaten im Mittel ein besseres Ergebnis erzielen, wenn das Thema des Textes eine Nähe zum Studienziel aufweist (Tabelle 42, Seite 271). Es wurden zwei Mittelwerte der Diagonalen berechnet: für die "Experten" und für die "Ahnungslosen". In die Berechnung des Mittelwerts für die "Experten" fließen folgende Werte ein: die Leistungen der Kandidaten aus den T-Kursen im Leseverstehenstest "Geschwindigkeitsmessung" sowie die Leistungen der Kandidaten aus den W-Kursen im Leseverstehenstest "Inflation". Der Ausdruck "Experte" ist für die Kollegiaten möglicherweise unzutreffend; er drückt den Sachverhalt aber plastisch aus. In die Berechnung des Mittelwerts für die "Ahnungslosen" fließen folgende Daten ein: die Ergebnisse der Kollegiaten aus den T-Kursen im Leseverstehenstest "Inflation" sowie die Ergebnisse der Kollegiaten aus den W-Kursen im Leseverstehenstest "Geschwindigkeit". Auch dieser Ausdruck dient in erster Linie zur Veranschaulichung der Situation, welche durch den Mittelwert ausgedrückt wird. Die Mittelwerte der Diagonalen zeigen also das Zusammenwirken zwischen Gruppen und Leseverstehenstests mit Fachbezug und geben damit Auskunft über den Effekt des Fachbezugs auf die Testleistungen. Die "Experten" erzielten in dem Leseverstehenstest mit Bezug zu ihrem Studienfach 60 Prozent, die "Ahnungslosen" erzielten in dem Test mit fremdem Thema im Mittel lediglich 48 Prozent.

Zusammenfassung und Diskussion: Bei einer Unterscheidung der Kandidaten nach ihrem Studienziel und der Kurszugehörigkeit im Studienkolleg – und diese Unterscheidung liegt der Kurszuweisung zugrunde – erzielen Kandidaten im Mittel ein signifikant besseres Ergebnis, wenn das Thema des Textes eine Nähe zum zukünftigen Studienfach hat. Welche Ursache gibt es für die höheren Leistungen, welche Kandidaten im Leseverstehenstest aus ihrem zukünftigen Studienfach erzielten? Nahe liegend ist, dass Vorkenntnisse zum Thema eine Rolle spielen. In Kapitel 6.1.4 (Seite 260 f) wurde zwar darauf hingewiesen, dass man nicht davon ausgehen kann, dass KURS in jedem Fall Vorkenntnisse der Kandidaten zum Thema erfasst. Dennoch liegt diese Interpretation hier nahe: Möglich ist, dass die Kandidaten aufgrund von Vorkenntnissen in Tests aus dem zukünftigen Studienfach bessere Ergebnisse erzielten. Man könnte den Studienwunsch mit einiger Berechtigung als Indikator für ein besonderes Interesse an bestimmten Themen ansehen. Das zu vermutende Interesse könnte dazu führen, dass bessere Leistungen in Tests mit Bezug zum Studienfach erzielt werden. Hinzu kommt,

dass alle Testteilnehmer im Rahmen der Studienkollegs bereits an Fachunterricht zu wirtschaftlichen Themen oder zu technischen Themen teilnahmen.

Es fällt auf, dass Kandidaten aus W-Kursen ein signifikant höheres Ergebnis in C-TEST erzielen, was ich als Ausdruck einer höheren (deutschen) Sprachkompetenz werte. Das heißt, KURS erfasst nicht Vorkenntnisse allein, sondern zu einem gewissen Teil auch Unterschiede in der Deutschkompetenz. Dieser Zusammenhang dürfte dazu geführt haben, dass der Unterschied zwischen den Gruppen beim Leseverstehenstest INFLATION über 13 Prozent beträgt, beim Leseverstehenstest GESCHWINDIGKEIT aber nur 7,7 Prozent.

Tabelle 41: Ergebnisse nach Kurszugehörigkeit/Studienziel (Variable KURS) – Mittelwerte und Signifikanz der Unterschiede zwischen den Mittelwerten

Kurszugehörigkeit	**Mittelwert im LV-Inflation (*AM*) in %**	**Mittelwert im LV-Geschwindigkeit (*AM*) in %**
Gesamt	54,0 % n = 352	55,9 % n = 333
T-Kurs Technische Studiengänge	48,1 % n = 193	59,2 % n = 192
W-Kurs Wirtschaftliche Studiengänge	61,3 % n = 159	51,5 % n = 141
Differenz der Mittelwerte und Signifikanzniveau	**13,2 %** F = 40,347; $p < 0,01$	**7,7 %** F = 9,830; $p < 0,01$

Anm.: Ergebnisse einfaktorieller Varianzanalysen.

Tabelle 42: Ergebnisse nach Kurszugehörigkeit/Studienziel (Variable KURS) – diagonaler Mittelwert

Diagonaler Mittelwert: die "Experten" (W-Kurs/Inflation, T-Kurs/Geschwindigkeit)	**Diagonaler Mittelwert: die "Ahnungslosen"** (T-Kurs/Inflation, W-Kurs/Geschwindigkeit)
60 % n = 335	48 % n = 321

Tabelle 43: Ergebnisse im C-Test nach Kurszuweisung/Studienziel (Variable KURS) – Mittelwerte und Signifikanzniveaus

	Kandidaten aus Wirtschaftskursen	**Kandidaten aus Technikkursen**	**Differenz**	**Signifikanz**
Ergebnisse im C-Test (*AM*)	56,7 % n = 168	51,1 % n = 202	5,6 %	F = 6,537; $p = 0,011$

Vorkenntnisse nach BEGRIFFE

Ergebnisse: Vor dem Lesen wurden die Kandidaten zu ihren Vorkenntnissen zu den jeweiligen Themen befragt. Waren sie mit den Schlüsselbegriffen aus dem Text vertraut, so wurden Vorkenntnisse angenommen. Die Ergebnisse der nach Vorkenntnissen eingeteilten Gruppen in den Leseverstehenstests gehen aus der Tabelle 44 hervor.

Wenn sich die Kandidaten vor dem Lesen der Texte zu zentralen Schlüsselbegriffen äußern konnten, erzielten sie hohe Ergebnisse: Im Leseverstehenstest "Inflation" 58 Prozent und im Test "Geschwindigkeit" 61 Prozent. Wenn sie mit den Schlüsselbegriffen nicht vertraut waren, lagen die Ergebnisse deutlich darunter: Der Mittelwert im Leseverstehenstest "Inflation" war mit 44 Prozent um 15 Prozent niedriger. Im Test "Geschwindigkeit" waren die Mittelwerte mit 46 Prozent um 15 Prozent geringer. Der Unterschied ist hoch signifikant.

Zusammenfassung und Diskussion: Diese Beobachtungen deuten darauf hin, dass Vorkenntnisse auch dann eine deutliche Rolle für das Verständnis spielen, wenn der Fachlichkeitsgrad der Texte gering ist. Diese Unterschiede sind enorm. Gibt es möglicherweise noch andere Ursachen für die Leistungsunterschiede? Es ist denkbar, dass auch von der Variablen BEGRIFFE Deutschkenntnisse erfasst werden, da die Kandidaten die Schlüsselbegriffe in der Zielsprache (d. h. auf Deutsch) erläutern mussten. Es ist denkbar, dass Kandidaten mit fortgeschrittenen Deutschkenntnissen die Fragen zu den Schlüsselbegriffen eher positiv beantworten und eher einen Antwortversuch unternehmen als Kandidaten mit weniger fortgeschrittenen Deutschkenntnissen. Dann wären die Mittelwertunterschiede nicht nur dem Einfluss der Vorkenntnisse zuzuschreiben, sondern auch dem Einfluss der Deutschkompetenz. Dies kann gezeigt werden, wenn man die Ergebnisse der C-Tests hinzuzieht. Tabelle 45 zeigt die Ergebnisse der Kandidaten im C-Test (und nicht in den Leseverstehenstests). Die Einteilung der Kandidaten geschah nach Kenntnis der Schlüsselbegriffe vor dem Lesen (Variable BEKANNT). Es stellt sich heraus, dass die Kandidaten mit Vorkenntnissen in den Leseverstehenstests höhere Ergebnisse im C-Test erzielten als die Kandidaten ohne Vorkenntnisse. Die Unterschiede sind jeweils hoch signifikant (Tabelle 45).

Wenn man die Ergebnisse der Kandidaten in Gruppen betrachtet, die nach Vorkenntnissen gebildet werden, ergibt sich folgendes Bild: Bei den Leseverstehenstests "Inflation"

und "Geschwindigkeit" erzielten die Kandidaten mit Vorkenntnissen ein deutlich höheres Ergebnis als die Kandidaten ohne Vorkenntnisse. Man muss allerdings davon ausgehen, dass die Kandidaten mit Vorkenntnissen (nach BEGRIFFE) im Mittel auch über bessere Deutschkenntnisse verfügten. Für die Ergebnisunterschiede könnten daher auch unterschiedliche Deutschkenntnisse verantwortlich sein.

Tabelle 44: Ergebnisse in Leseverstehenstests nach Kenntnis der Schlüsselbegriffe vor dem Lesen (Variable BEGRIFFE) – Mittelwerte

	Test	**ja**	**nein**	**Differenz und Signifikanz**	**gesamt**
Begriffe "Inflation" oder "Deflation" bekannt	Ergebnisse im LV-Inflation (*AM*)	58,4 % n = 307	44,0 % n = 198	14,5 % F = 68,457; $p < 0{,}01$	52,8 % n = 505
Begriffe "Radar" oder "Laser" bekannt	Ergebnisse im LV-Geschwindigkeit (*AM*)	61,3 % n = 199	46,8 % n = 287	14,5 % F = 53,180; $p < 0{,}01$	52,8 % n = 486

Anm.: Ergebnisse einfaktorieller Varianzanalysen.

Tabelle 45: Ergebnisse im C-Test nach Vorkenntnissen zu Fachthemen (Variable BEGRIFFE) – Mittelwerte und Signifikanzniveaus

	Test	**ja**	**nein**	**Differenz**	**Signifikanz**
Vorkenntnisse zum Text "Inflation"	Ergebnisse im C-Test (*AM*)	58,6 % n = 308	51,5 % n = 203	7,1 %	F = 17,867; $p < 0{,}01$
Vorkenntnisse zum Text "Geschwindigkeit"	Ergebnisse im C-Test (*AM*)	61,4 % n = 201	51,3 % n = 302	10,1	F = 36,777; $p < 0{,}01$

Vorkenntnisse nach BEKANNT

Ergebnisse: In diesem Abschnitt betrachte ich die Ergebnisse unter der Fragestellung, ob die Kandidaten den Text (nach dem Lesen) als bekannt oder als unbekannt einstuften (Variable BEKANNT). Die Kandidaten, welche mit den Themen der Texte bereits vertraut waren, erzielten im Mittel ein höheres Ergebnis als die Kandidaten, die nach dem Lesen angaben, das Thema sei ihnen "eher unbekannt". Beim Leseverstehenstest "Geschwindigkeit" beträgt der Unterschied zwölf Prozent, bei "Inflation" dreizehn Prozent. Diese Unterschiede sind jeweils hoch signifikant (Tabelle 46).

Zusammenfassung und Diskussion: Ob es an dem Einfluss der Vorkenntnisse liegt oder ob die Unterschiede zwischen den Gruppen den unterschiedlichen Sprachkenntnissen zugeschrieben werden sollten, kann mit den Ergebnissen im C-Test erläutert werden (siehe Tabelle 48). Die Gruppe der Kandidaten, denen das Thema "Inflation" bereits vertraut war, erreichten im C-Test höhere Ergebnisse als diejenigen, die mit dem Thema noch nicht vertraut waren. Die Differenz der Mittelwerte ist zwar auf dem Niveau 95 Prozent signifikant, sie beträgt aber nur drei Prozent. Die Ergebnisse der 231 Kandidaten, die mit dem Thema "Geschwindigkeit" bereits vertraut waren, unterscheiden sich nicht signifikant von den Ergebnissen der 158 Kandidaten, denen das Thema neu war. Anders als mit den Variablen KURS und BEGRIFFE ist die Variable BEKANNT somit als relativ unabhängig von dem Niveau der Deutschkenntnisse anzusehen.

Es ist anzunehmen, dass die unterschiedlichen Ergebnisse in den Lebensverstehenstests nicht auf dem Einfluss unterschiedlicher Sprachkenntnisse beruhen, sondern vor allem dem Einfluss der unterschiedlichen Vorkenntnisse geschuldet sind. Es wurde bei der Vorstellung der Variablen zu den Vorkenntnissen in Kapitel 6.1.4 bereits darauf hingewiesen, dass die Variable BEKANNT nicht oder nur in sehr geringem Umfang mit C-TEST korreliert, demnach kein Indikator für Deutschkenntnisse darstellt.

Die Abbildung 31 (Seite 276) und die Abbildung 32 (Seite 277) visualisieren die unterschiedlichen Leistungen in Leseverstehenstests nach Vorkenntnissen. In der Zusammenfassung wird deutlich, dass Gruppen mit Vorkenntnissen (Säulen auf der linken Seite) höhere Ergebnisse erzielten als Gruppen ohne Vorkenntnisse (Säulen auf der rechten

Seite). Die Unterschiede sind groß, und sie sind treten relativ unabhängig von der Erhebungsmethode auf.

Tabelle 46: Ergebnisse in Leseverstehenstests nach Vorkenntnissen laut Selbstauskunft nach dem Lesen (Variable BEKANNT) – Mittelwerte und Signifikanzniveaus

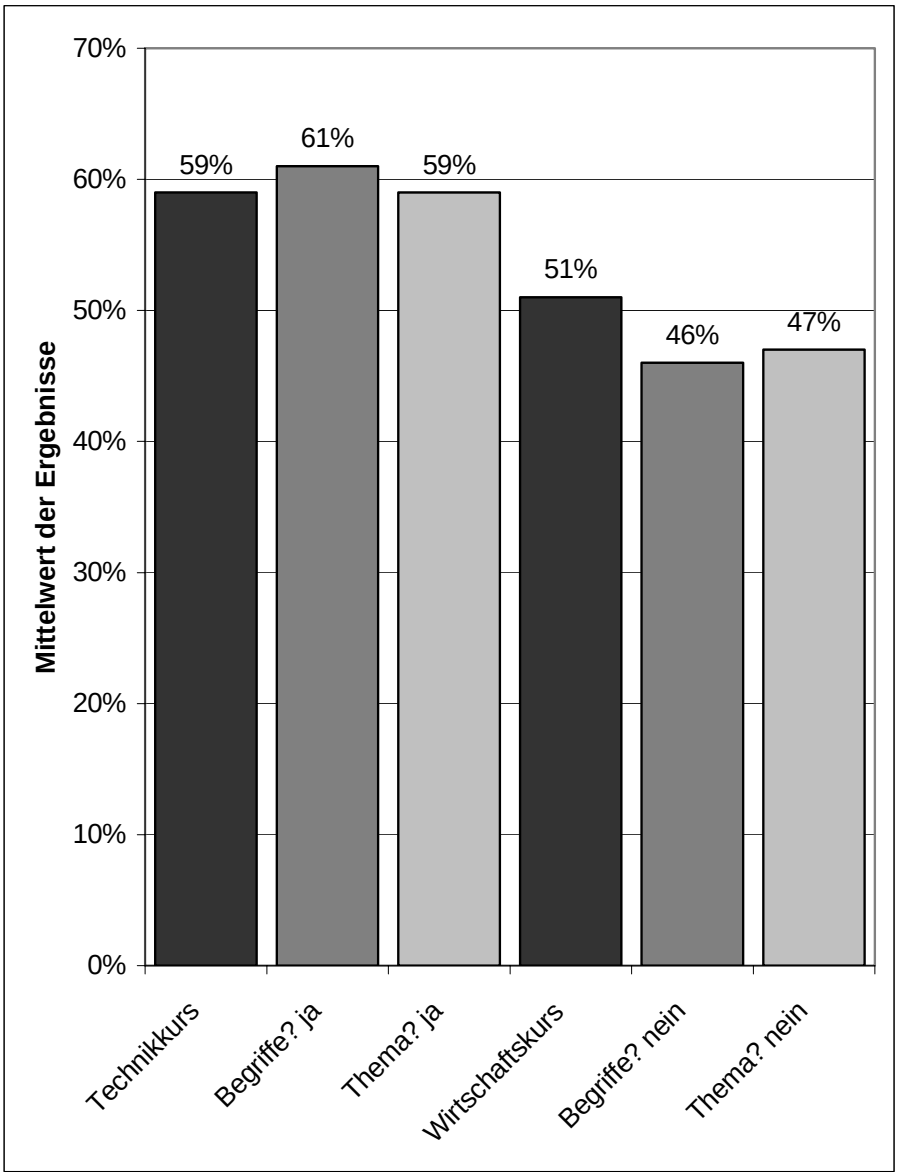
	ja	**nein**	**gesamt**	**Differenz**	**Signifikanz**
Text "Inflation" bekannt?	57,9 % n = 314	44,6 % n = 196	52,8 % n = 510	13,3 %	F = 57,003; $p < 0{,}01$
Text "Geschwindigkeit" bekannt?	59,4 % n = 230	47,2 % n = 259	52,9 % n = 489	12,2 %	F = 38,263; $p < 0{,}01$

Tabelle 47: Ergebnisse im C-Test nach Vorkenntnissen zu Fachthemen (Variable BEGRIFFE) – Mittelwerte und Signifikanzniveaus

	Test	**ja**	**nein**	**Differenz**	**Signifikanz**
Vorkenntnisse zum Text "Inflation"	Ergebnisse im C-Test (*AM*)	58,6 % n = 308	51,5 % n = 203	7,1 %	F = 17,867; $p < 0{,}01$
Vorkenntnisse zum Text "Geschwindigkeit"	Ergebnisse im C-Test (*AM*)	61,4 % n = 201	51,3 % n = 302	10,1	F = 36,777; $p < 0{,}01$

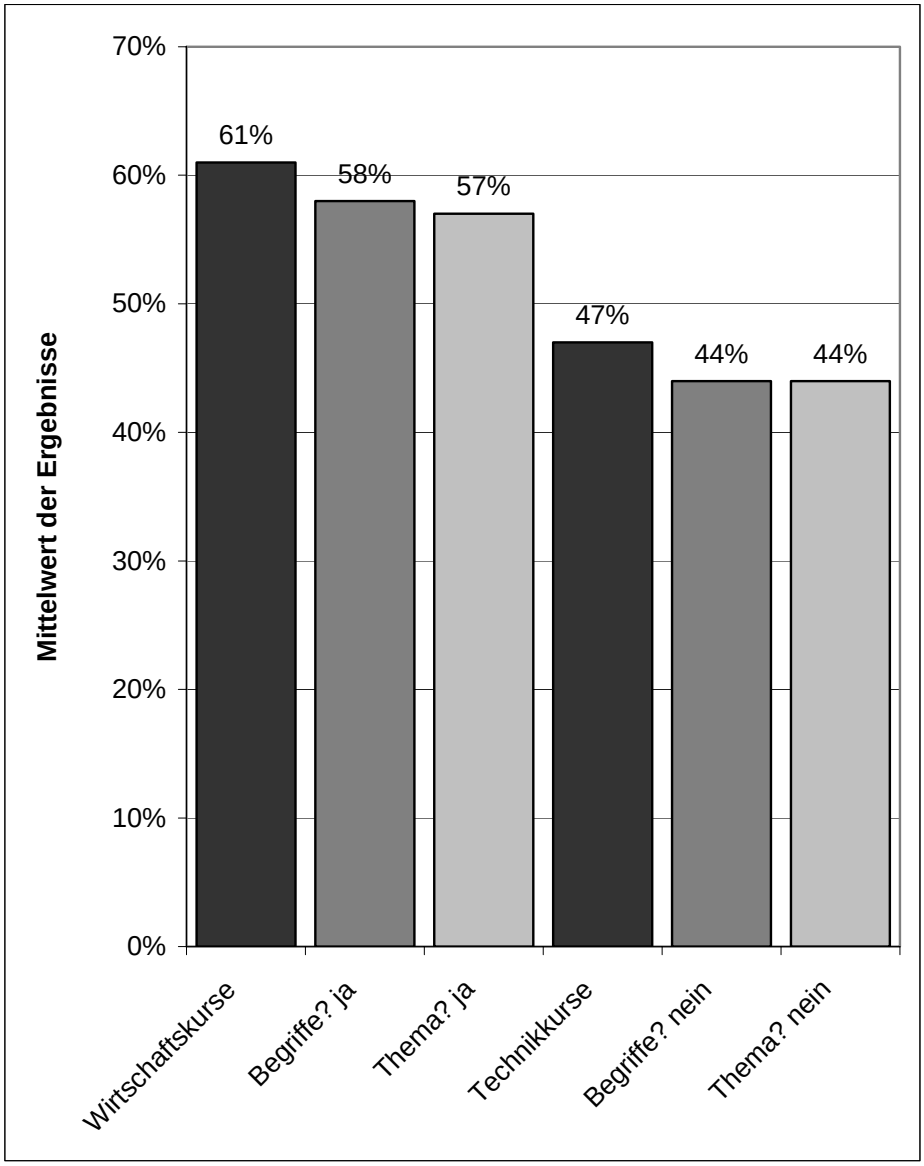
Tabelle 48: Ergebnisse im C-Test nach Vorkenntnissen zu Fachthemen (Variable BEKANNT) – Mittelwerte und Signifikanzniveaus

	Test	**ja**	**nein**	**gesamt**	**Differenz**	**Signifikanz**
Text "Inflation" bekannt?	Ergebnisse im C-Test (*AM*)	57,1 % n = 312	53,7 % n = 197	55,8 % n = 509	3,4 %	F = 3,849 $p = 0{,}05$
Text "Geschwindigkeit" bekannt?	Ergebnisse im C-Test (*AM*)	54,6 % n = 231	56,0 % n = 258	55,3 % n = 489	1,6 %	F = 0,675; nicht signifikant



Anm.: Anzahl (n) siehe Tabelle 36, Tabelle 37 und Tabelle 38 (Seite 258, 259 und 261).

Abbildung 31: Ergebnisse im Leseverstehenstest "Geschwindigkeit" nach Vorkenntnissen (Säulendiagramm)



Anm.: Anzahl (n) siehe Tabelle 36, Tabelle 37 und Tabelle 38 (Seite 258, 259 und 261).

Abbildung 32: Ergebnisse im Leseverstehenstest "Inflation" nach Vorkenntnissen (Säulendiagramm)

6.2.2. Vorkenntnisse oder Deutschkenntnisse?

Einfluss der Deutschkenntnisse

Ergebnisse: Bevor ich den Einfluss der Vorkenntnisse auf die Ergebnisse in den Leseverstehenstests in Abhängigkeit von den Deutschkenntnissen betrachte, möchte ich kurz auf die Rolle der Deutschkenntnisse eingehen, ohne die Vorkenntnisse zu berücksichtigen. Gibt es einen Zusammenhang zwischen den Ergebnissen im C-Test und den Leistungen in den Leseverstehenstests?

Die Korrelationskoeffizienten nach Pearson zwischen den Ergebnissen im C-Test (als Hinweis auf das Niveau der Deutschkenntnisse) und den Leseverstehenstests "Geschwindigkeit" bzw. "Inflation" deuten auf einen mittleren Zusammenhang (r = 0,422 bzw. r = 0,457). Sowohl die Korrelationen des C-Tests mit dem Leseverstehen "Inflation" also auch mit dem Leseverstehenstest "Geschwindigkeit" sind hoch signifikant (Tabelle 49). Der Zusammenhang liegt leicht unter den Korrelationen der beiden Leseverstehenstests untereinander (r = 0,584). Ein Zusammenhang zwischen den Deutschkenntnissen und den Ergebnissen in Leseverstehenstests mit geringem Fachlichkeitsgrad ist erkennbar (siehe Abbildung 33 und Abbildung 34).

Tabelle 49: Ergebnisse in Leseverstehenstests und im C-Test – Korrelationen nach Pearson

	C-Test	**Leseverstehen Geschwindigkeit**	**Leseverstehen Inflation**
C-Test	_	,422 n = 488, $p < 0,01$	,457 n = 509, $p < 0,01$
Leseverstehen Geschwindigkeit	,422 n = 488, $p < 0,01$	_	,584 n = 486, $p < 0,01$
Leseverstehen Inflation	,457 n = 509, $p < 0,01$	,584 n = 486, $p < 0,01$	_

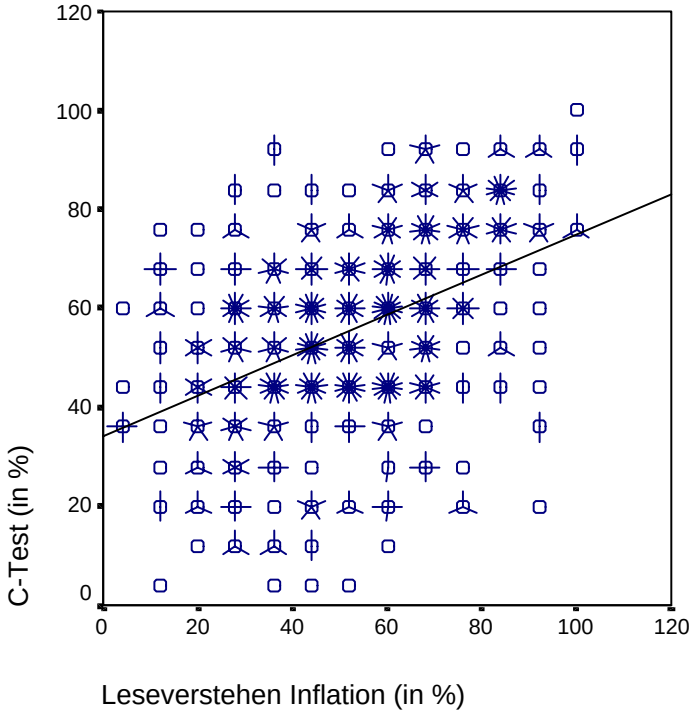


Abbildung 33: Ergebnisse im C-Test und im Leseverstehenstest "Inflation" – Sonnenblumen-Streudiagramm mit Regressionsgeraden (n = 509)

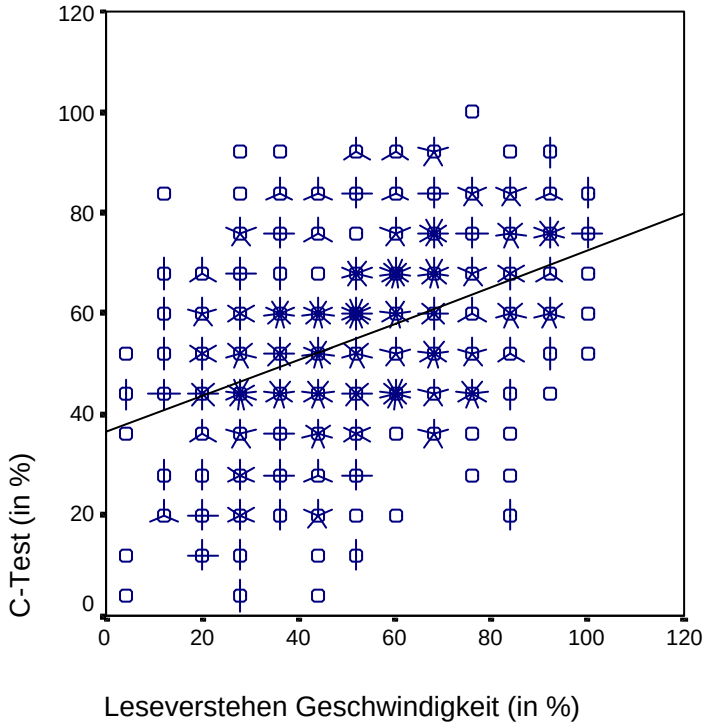


Abbildung 34: Ergebnisse im C-Test und im Leseverstehenstest "Geschwindigkeit" (Sonnenblumen-Streudiagramm mit Regressionsgeraden; n = 501)

Dass das Niveau der Deutschkenntnisse einen Einfluss auf die Leistungen in Leseverstehenstests hat, lässt sich auch anhand von einfaktoriellen Varianzanalysen zeigen. Dabei wurden die Leistungen im C-Test als Kovariate und die Ergebnisse in den Leseverstehenstests "Inflation" bzw. "Geschwindigkeit" als abhängige Variablen behandelt. Es ergibt sich ein hoch signifikanter Einfluss der Ergebnisse im C-Test auf die Leistungen in den Leseverstehenstests (LV "Inflation": $F = 127{,}838$; $p < 0{,}01$; $n = 509$; LV "Geschwindigkeit": $F = 112{,}803$; $p < 0{,}01$; $n = 488$). Diese Unterschiede gehen auch aus dem Boxplot-Diagramm hervor, bei dem Ergebnisse im Leseverstehenstest "Geschwindigkeit" und "Inflation" nach Leistungen im C-Test dargestellt sind (Abbildung 35 und Abbildung 36).

Zusammenfassung und Diskussion: In dieser Studie waren die Ergebnisse in den beiden Leseverstehenstests ähnlicher als die Ergebnisse im C-Test und den jeweiligen Leseverstehenstests. Neben den Vorkenntnissen, deren Einfluss auf das Leseverstehen mit geringem Fachlichkeitsgrad in Kapitel 6.2.1 (Seite 268 ff) aufgezeigt wurde, spielt auch das Niveau der Deutschkenntnisse eine signifikante Rolle für die Leistungen in Leseverstehenstests. Welche Variable einen entscheidenderen Einfluss hat, wird im folgenden Abschnitt erläutert.

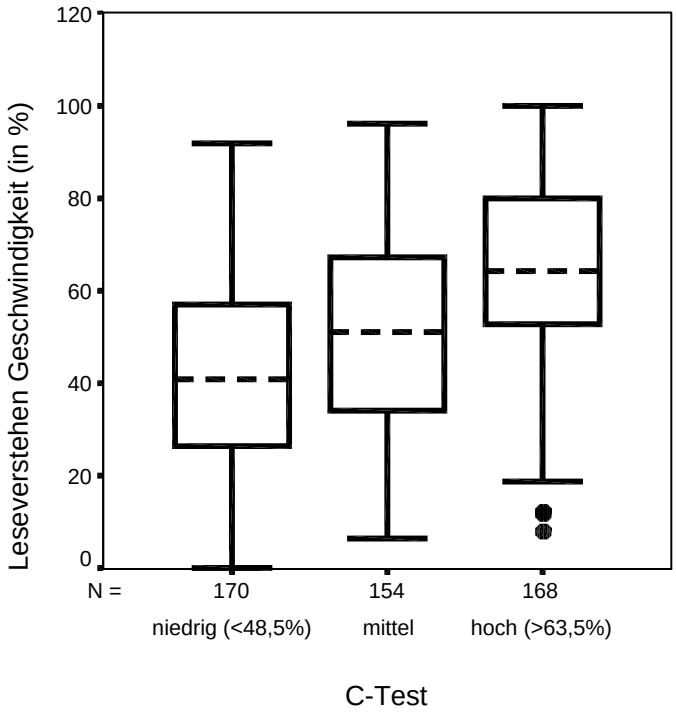


Abbildung 35: Ergebnisse im Leseverstehenstest "Geschwindigkeit" nach Leistungen im C-Test (Boxplot mit Median, Interquartilbereich, Ausreißern und Extremfällen)

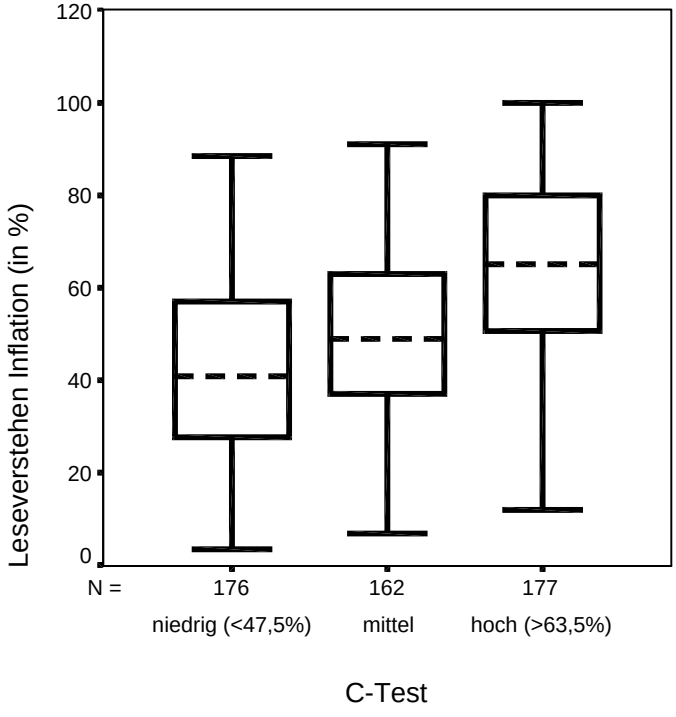


Abbildung 36: Ergebnisse im Leseverstehenstest "Inflation" nach Leistungen im C-Test (Boxplot mit Median, Interquartilbereich und Extremfällen)

Korrelationen und Regressionsanalysen

Ergebnisse: Mit multiplen Regressionsanalysen soll festgestellt werden, welche Variablen einen Einfluss auf die Leistungen der Kandidaten in Leseverstehenstests mit geringem Fachlichkeitsgrad aufweisen. Von Interesse ist, wie stark der jeweilige Einfluss ist. Zu den Einflussvariablen, die zur Verfügung stehen, zählen nicht nur die Vorkenntnisse, welche mit drei unterschiedlichen Methoden erhoben wurden (Kapitel 6.1.4, Seite 257), sondern auch die Sprachkenntnisse, welche mit Hilfe des C-Tests erhoben wurden (Kapitel 6.1.3, Seite 255). Als abhängige Variablen werden die Ergebnisse in den beiden Leseverstehenstests mit geringem Fachlichkeitsgrad (INFLATION bzw. GESCHWINDIGKEIT) in die Regressionsanalyse eingegeben. Die unabhängigen Variablen sind:

- Niveau der Deutschkenntnisse laut C-Test (C-TEST),
- Vorkenntnisse nach Kurszuordnung/Studienziel (KURS),
- Vorkenntnisse nach Kenntnis der Schlüsselbegriffe vor dem Lesen (BEGRIFFE),
- Vorkenntnisse laut Selbstauskunft nach dem Lesen (BEKANNT).

Einige Anmerkungen zur Codierung: Die Regressionsanalyse erfordert bei nominalskalierten Daten die Eingabe in dichotomer Form. Daher wurden bei der Variablen KURS die Kandidaten ausgeschlossen, die nicht in T- bzw. W-Kursen sind. Bei der Regressionsanalyse mit INFLATION wurden die Kandidaten aus W-Kursen mit 1 codiert und die Kandidaten aus T-Kursen mit 0, weil das Studienziel "Wirtschaft" bzw. die Kurszuweisung "W-Kurs" als Hinweis auf Vorkenntnisse zum Thema "Inflation" interpretiert wird. Bei der Regressionsanalyse mit GESCHWINDIGKEIT wurde "T-Kurs" mit 1 und "W-Kurs" mit 0 codiert. Die Vorkenntnisse nach Kenntnis der Schlüsselbegriffe vor dem Lesen (Variable BEGRIFFE) wurden wie folgt codiert: Waren die Begriffe bekannt, so wurde mit 1 codiert, waren sie nicht bekannt, mit 0. Ebenso wurde bei den Vorkenntnissen laut Selbstauskunft nach dem Lesen (BEKANNT) verfahren. Bei den Daten zu den Vorkenntnissen gehe ich von einem ordinalen Messniveau aus. Die Daten zum Studienziel bzw. zur Kurszuweisung beruhen auf einer nominalen Unterscheidung. In Kapitel 6.2.1 (Seite 268 ff) konnte jedoch gezeigt werden, dass die Variable KURS einen signifikanten Einfluss auf das Leseverstehen hat. Sie spiegelt eine Ordnung der Kandidaten und wird daher dem ordinalen Messniveau zugeschrieben. Die Deutschkenntnisse nach C-Test (C-TEST) wurden mit

den Ergebnissen aus dem C-Test auf einer Skala von Null bis 100 eingegeben, welche auf den Prozentwerten beruht.

Ich beginne mit der Korrelation der abhängigen Variablen und INFLATION. Die Korrelationsanalysen wurden mit den Ergebnissen der Kandidaten durchgeführt, die von allen Variablen erfasst wurden (listenweiser Fallauschluss). Die Korrelationen der unabhängigen Variablen untereinander sind unterschiedlich hoch: Ein mittlerer Zusammenhang war zwischen BEGRIFF und BEKANNT zu beobachten, sowie zwischen BEGRIFF und KURS. Zusammenhänge zwischen Variablen zu den Vorkenntnissen ergeben sich aus der Nähe der gemessenen Konstrukte. Wenn Kandidaten die Schlüsselbegriffe aus dem Text kennen, ist es wahrscheinlich, dass der Inhalt der Texte für sie bereits bekannt ist. Die Zusammenhänge zwischen den Variablen zu den Vorkenntnissen sind also beabsichtigt, bei der Regressionsanalyse allerdings störend, da sich der Einfluss einer einzelnen Variable nicht von den anderen abgrenzen lässt. Im vorliegenden Fall bleibt das folgenlos, da es in erster Linie um den Vergleich der Vorkenntnisse mit den Sprachkenntnissen geht, nicht aber um einen Vergleich der einzelnen Variablen zu den Vorkenntnissen untereinander.

Die Ergebnisse aus dem Leseverstehenstest "Inflation" korrelierten am höchsten mit den Ergebnissen aus dem C-Test (Korrelationskoeffizient nach Spearman: $r_s = 0{,}490$; $p < 0{,}01$). Die Korrelationen mit den übrigen Variablen zu den Vorkenntnissen sind etwas niedriger, aber ebenfalls hoch signifikant (Tabelle 50). Von Interesse ist weiter, dass das Niveau der Deutschkenntnisse (C-TEST) mit den Vorkenntnissen (KURS, BEGRIFFE und BEKANNT) korreliert. Die Korrelationen liegen im niedrigen Bereich und sind auf dem Niveau 95 bzw. 99 Prozent signifikant. Dies ist ein weiterer Hinweis darauf, dass die Variablen BEGRIFFE und BEKANNT nicht nur Vorkenntnisse, sondern auch Deutschkenntnissen erheben. Dieser Aspekt muss bei der Interpretation berücksichtigt werden.

Tabelle 50: Sprachkenntnisse, Vorkenntnisse und Leseverstehen "Inflation" – Korrelationen (n = 341)

	C-TEST	KURS	BEGRIFFE	BEKANNT	INFLATION
C-TEST	_	0,136**	0,194**	0,112*	0,345**
KURS	0,165*	_	0,505**	0,322**	0,285**
BEGRIFFE	0,235**	0,505**	_	0,578**	0,350**
BEKANNT	0,136*	0,322**	0,578**	_	0,314**
INFLATION	0,490**	0,348**	0,427**	0,383**	_

Anm: Korrelationskoeffizienten in der obere Dreiecksmatrix gemäß Kendall-Tau-b, in der unteren Spearman-rho.
** = Korrelationen sind auf dem Niveau 99 % signifikant.
* = Korrelationen sind auf dem Niveau 95 % signifikant.

Tabelle 51: Leseverstehen "Inflation" – Ergebnisse der Regressionsanalyse (n = 341)

Variablen	Nicht standardisierter Koeffizient B	standardisierter Koeffizient Beta	T	Signifikanz
Konstante	19,089		7,147	p < 0,01
C-TEST	0,425	0,413	9,428	p < 0,01
BEKANNT	8,312	0,193	3,692	p < 0,01
BEGRIFFE	7,054	0,165	2,847	p < 0,01
KURS	5,046	0,122	2,456	p < 0,05

	Korrigiertes R-Quadrat
C-TEST	0,243
C-TEST und BEGRIFFE	0,353
C-TEST, BEGRIFFE und BEKANNT	0,379
C-TEST, BEGRIFFE, BEKANNT und KURS	0,390

Bei der Regressionsanalyse mit dem Leseverstehenstest "Inflation" als abhängige Variable wurde die "schrittweise Methode" verwendet. Es sollten die Koeffizienten des folgenden Modells geschätzt werden:

Leseverstehenstest =
Konstante + b1 * C-TEST + b2 * KURS + b3 * BEGRIFFE + b4 * BEKANNT

Mit den Ergebnissen von 341 Kandidaten konnte ein signifikantes Modell ermittelt werden, welches alle vier Prädiktorvariablen einschloss. Die Regressionsgleichung zur Vorhersage des Ergebnisses im Leseverstehenstest "Inflation" lautet bei Einschluss aller vier unabhängigen Variablen (siehe Tabelle 51):

INFLATION =
19,089 + 0,425*C-TEST + 8,312*BEKANNT + 7,054*BEGRIFFE + 5,046*KURS

Die Regressionsgleichung stellt folgenden Zusammenhang dar: Wenn man das Ergebnis im Leseverstehenstest "Inflation" vorhersagen möchte, sollen zur Konstante 19,089 noch folgende Terme addiert werden: das Ergebnis im C-Test, welches man mit 0,425 multipliziert; wenn die Schlüsselbegriffe aus dem Text bekannt waren, sollen 7,054 addiert werden; wenn das Thema laut Selbstauskunft nach dem Lesen bereits bekannt war, addiert man 8,312; wenn der Kandidat ein wirtschaftswissenschaftliches Studium anstrebt, addiert man schließlich 5,046.

Verständlich wird die Bedeutung der Variable C-TEST auch am standardisierten Koeffizienten Beta, der den Einfluss der Variablen unabhängig vom Skalenumfang darstellt. Allerdings können nicht die Werte aus Tabelle 51 verwendet werden, sondern es sollte der Einfluss von C-TEST und einzelnen Variablen zu Vorkenntnissen verglichen werden. Wenn in die Regressionsanalyse nur C-TEST und BEGRIFFE einbezogen werden, liegt der standardisierte Koeffizient Beta bei 0,419 für C-TEST und 0,337 für BEGRIFFE.

Wie gut kann man das Ergebnis mithilfe der Regressionsfunktion vorhersagen? An dem Bestimmtheitsmaß, dem korrigierten "R-Quadrat", kann man erkennen, wie gut die Regressionsfunktion die tatsächlichen Ergebnisse abbildet. Unter Einschluss aller vier unabhängigen Variablen liegt das korrigierte R-Quadrat bei 0,390. Das bedeutet, dass die vier Variablen statistisch 39 Prozent der Varianz der Ergebnisse erklären. Wenn man Variablen aus der Regressionsfunktion ausschließt, sinkt dieser Wert nur leicht:

Lässt man die Variable KURS heraus, erklärt die Gleichung 38 Prozent der Ergebnisvarianz ($r^2_{KORR} = 0{,}379$). Erst wenn alle Variablen zu den Vorkenntnissen ausgeschlossen werden und nur C-TEST beibehalten wird, sinkt der Wert auf 25 Prozent ($r^2_{KORR} = 0{,}246$).

Eine weitere Regressionsanalyse wurde mit dem Leseverstehenstest "Geschwindigkeit" als abhängiger Variable durchgeführt. Die unabhängigen Variablen waren wiederum die Sprachkenntnisse mittels der Ergebnisse im C-Test sowie die drei Variablen zu den Vorkenntnissen, KURS, BEGRIFFE und BEKANNT. Mit Ausnahme der Variable C-TEST sind es andere Variablen als bei der vorangegangenen Regressionsanalyse mit dem Leseverstehenstest "Inflation", da sich die Vorkenntnisse nun auf den Text "Geschwindigkeit" beziehen. Auch die Variable KURS wurde umcodiert: Kandidaten aus den T-Kursen wurde der Wert 1 und Kandidaten aus W-Kursen der Wert 0 zugewiesen.

Zunächst stelle ich die Korrelationen zwischen den Variablen vor, die mit den Rangkorrelationskoeffizienten Spearman-rho bzw. Kendall-Tau-b bestimmt wurden (Tabelle 52). Der Zusammenhang zwischen den Variablen KURS, BEGRIFFE und BEKANNT, mit denen die Vorkenntnisse zum Text "Geschwindigkeit" erhoben wurden, ist gering. Keinen Zusammenhang gibt es zwischen dem Studienziel (KURS) und der Kenntnis der Schlüsselbegriffe (BEGRIFFE). Auch die Zusammenhänge zwischen den Deutschkenntnissen und den Variablen zu den Vorkenntnissen sind geringer als beim Leseverstehenstest "Inflation". Der negative Korrelationskoeffizient zwischen den Variablen KURS und C-TEST ist mit der Codierung zu erklären. Kandidaten aus W-Kursen erzielten ein signifikant höheres Ergebnis im C-Test als Kandidaten aus T-Kursen. Bei der Codierung 1 für "T-Kurs" und 0 für "W-Kurs" kommt es folglich zu einem negativen Wert.

Die Ergebnisse im Leseverstehenstest "Geschwindigkeit" korrelieren niedrig, aber hoch signifikant mit den Variablen zu den Vorkenntnissen.

Tabelle 52: Sprachkenntnisse, Vorkenntnisse und Leseverstehen "Geschwindigkeit" – Korrelationen (n = 324)

	C-TEST	KURS	BEGRIFFE	BEKANNT	GESCHWIN-DIGKEIT
C-TEST	_	-0,123**	0,245**	-0,074	0,321**
KURS	-0,149**	_	0,013	0,335**	0,123**
BEGRIFFE	0,296**	0,013	_	0,333**	0,266**
BEKANNT	-0,090	0,335**	0,333**	_	0,152**
GESCHWIN-DIGKEIT	0,458**	0,150**	0,324**	0,186**	_

** Korrelation ist auf dem Niveau 99 % signifikant (2-seitig).
Anm: Korrelationskoeffizienten in der oberen Dreiecksmatrix: Kendall-Tau-b; in der unteren: Spearman-rho.

Tabelle 53: Leseverstehen "Geschwindigkeit" – Ergebnisse der Regressionsanalyse (n = 325)

Variablen	nicht standardisierter Koeffizient B	Standardisierter Koeffizient Beta	T	Signifikanz
Konstante	18,375		3,474	p < 0,01
C-TEST	0,504	0,455	9,039	p < 0,01
KURS	7,942	0,179	3,548	p < 0,01
BEGRIFFE	6,191	0,139	2,610	p < 0,01
BEKANNT	5,478	0,123	2,309	p < 0,05

	Korrigiertes R-Quadrat
C-TEST	0,211
C-TEST und KURS	0,262
C-TEST, KURS und BEGRIFFE	0,285
C-TEST, KURS, BEGRIFFE und BEKANNT	0,297

Mit der Regressionsanalyse konnten signifikante Modelle entwickelt werden (siehe Tabelle 53). Der Einfluss aller vier Variablen ist auf dem Niveau 95 Prozent signifikant. Die Regressionsfunktion lautet:

$$\text{GESCHWINDIGKEIT} = 18{,}375 + 0{,}504*\text{C-TEST} + 7{,}942*\text{KURS} + 6{,}191*\text{BEGRIFFE} + 5{,}478*\text{BEKANNT}$$

Die Regressionsfunktion beschreibt, dass sich das Ergebnis im Leseverstehenstest "Geschwindigkeit" (in Prozent) vorhersagen lässt, wenn man zu 18,375 noch ungefähr die Hälfte des (prozentualen) Ergebnisses im C-Test addiert. Wenn die Kandidaten im T-Kurs sind bzw. technische Studiengänge anstreben, soll noch 7,942 addiert werden. Wenn die Kandidaten vor dem Lesen mit den Schlüsselbegriffen vertraut waren, soll 6,191 addiert werden. Wenn das Thema des Textes nach eigener Aussage bereits bekannt war, soll außerdem 5,478 hinzugefügt werden.

Im Vergleich zur Regressionsfunktion, die zum Leseverstehenstest "Inflation" ermittelt wurde, kann festgestellt werden: Wiederum können alle vier Variablen in ein signifikantes Modell einbezogen werden, wiederum ist der Einfluss der Sprachkenntnisse (C-Test) größer als der Einfluss der Variablen zu den Vorkenntnissen. Er ist im Falle des Leseverstehenstests "Geschwindigkeit" jedoch noch größer als beim Leseverstehenstest "Inflation". Dies spiegelt sich in dem höheren Koeffizienten B, der in diesem Fall 0,504 beträgt, während er bei INFLATION 0,390 beträgt. Die ermittelten Regressionsfunktionen unterscheiden sich durch ihre Aussagekraft. Während ein Modell unter Einschluss aller vier Variablen im Fall des Leseverstehenstest "Inflation" noch 39 Prozent der Variation erklären konnte, liegt dieser Wert bei dem Leseverstehenstest "Geschwindigkeit" unter 30 Prozent ($r^2_{KORR} = 0{,}297$). Wird nur die Variable C-Test in das Modell einbezogen, sinkt der Wert auf 21 Prozent ($r^2_{KORR} = 0{,}211$).

Zusammenfassung und Diskussion: Es konnte gezeigt werden, dass sowohl Vorkenntnisse als auch Deutschkenntnisse einen signifikanten Einfluss auf die Leistungen in den Leseverstehenstests mit Fachbezug haben. Als Ergebnis aus den Korrelations- und Regressionsanalysen kristallisieren sich drei Aspekte heraus:

- Die Bedeutung der Sprachkenntnisse bei den Leseverstehenstests mit Fachbezug scheint über derjenigen der Vorkenntnisse zu liegen. Die Variable C-TEST war bei den Regressionsanalysen jeweils der stärkste Prädiktor. Bei der Analyse mit

GESCHWINDIGKEIT lag sogar der kumulierte Wert des standardisierten Koeffizienten Beta aller Variablen zu den Vorkenntnissen unter dem Wert der Variablen C-TEST. Eigentlich sollte man aber die Beta-Werte von C-TEST nur mit jeweils einer Variablen zu den Vorkenntnissen vergleichen, da ähnliche Variablen nicht kumuliert werden sollten. Dann liegt der standardisierte Koeffizient Beta deutlich über den Werten der jeweils stärksten Variable zu den Vorkenntnissen. Möglicherweise ist der stärkere Einfluss der Deutschkenntnisse auch der Grund dafür, dass C-TEST stärker mit den Leseverstehenstests korreliert als die Variablen zu den Vorkenntnissen (siehe Tabelle 50 und Tabelle 52, Seite 284 und 287).

- Zweitens: Die Erklärung der Ergebnisse im Leseverstehenstest "Inflation" gelingt mit den Variablen zu den Sprachkenntnissen und zu den Vorkenntnissen nur zum Teil. Die ermittelten Modelle waren zwar signifikant, ihre Vorhersagekraft war jedoch gering. Fremdsprachenkenntnisse und Vorkenntnisse sind demnach nicht die einzigen Variablen, welche beim Leseverstehen in der Fremdsprache eine Rolle spielen. Dies ist nicht verwunderlich, denn das Konstrukt des Leseverstehens – auch bei Texten mit geringem Fachlichkeitsgrad – hängt nicht allein von Sprachkenntnissen und Vorkenntnissen ab. Zu den Variablen, die die Vorhersagekraft einer Regressionsfunktion erhöhen würden, könnten beispielsweise die Lesefähigkeit in der Muttersprache oder die Motivation zur Auseinandersetzung mit dem jeweiligen Text bzw. mit dem Test zählen. Es wurde bereits mehrfach darauf hingewiesen, dass Lesen ein komplexer Vorgang ist. Dies wird in der Regressionsanalyse bestätigt.

- Schließlich ist aus der Regressionsanalyse zu entnehmen, dass sich die Variablen zu den Vorkenntnissen nur leicht unterscheiden. Ihr kumulativer Beitrag zur Vorhersage der Ergebnisse im Leseverstehenstest "Inflation" ist zwar signifikant, insgesamt aber eher gering.

Mit den bisherigen Analysen konnten Informationen zur ersten Teilfrage des Kapitels gewonnen werden: Das Niveau der Deutschkenntnisse scheint eine größere Rolle für die Ergebnisse in Leseverstehenstests mit geringem Fachlichkeitsgrad zu spielen als Vorkenntnisse zum Thema. Ob der Einfluss der Vorkenntnisse zum Thema vom Niveau der Deutschkenntnisse abhängt, wird im folgenden Abschnitt mittels Varianzanalysen und Streudiagrammen ermittelt.

6.2.3. Vorkenntnisse und Deutschkenntnisse: Doppelte Schwellenhypothese

Die Streudiagramme in diesem Kapitel sind nach dem gleichen Muster aufgebaut: Die Ergebnisse einzelner Kandidaten in C-TEST sind auf der X-Achse, die Ergebnisse im Leseverstehenstest mit Fachbezug auf der Y-Achse abgebildet. Die Dreiecke stehen jeweils für Ergebnisse von Kandidaten ohne Vorkenntnisse, die Kreuze für Kandidaten mit Vorkenntnissen. Außerdem sind jeweils zwei Lowess-Regressionslinien abgebildet. Lowess-Anpassungslinien eignen sich besonders, um mögliche Änderungen der mittleren Tendenz aufzuzeigen, weil sie nur für Teile der Ergebnisse berechnet werden. Für jede Gruppe wurde eine eigene Lowess-Kurve abgebildet.

Der Verlauf der Lowess-Regressionslinien veranschaulicht folgende Aspekte:

- Horizontaler Verlauf oder Steigung? Ein horizontaler Verlauf würde darauf hindeuten, dass Deutschkenntnisse (nach C-TEST) die Ergebnisse in den Leseverstehenstests nicht beeinflussen. Ein sehr steiler Verlauf deutet auf einen starken Einfluss der Deutschkenntnisse hin.

- Auf einer Linie oder Abstand zwischen den Linien? Wenn beide Kurven mehr oder weniger auf einer Linie verlaufen würden, wäre der Einfluss der Vorkenntnisse zu vernachlässigen. Je größer der Abstand zwischen den Kurven, desto größer ist der Einfluss der Vorkenntnisse auf die Ergebnisse in dem jeweiligen Leseverstehenstest.

- Paralleler Verlauf oder nicht-paralleler Verlauf? Wenn es keine Abhängigkeit der Vorkenntnisse von den Sprachkenntnissen geben würde, würden die Kurven einen parallelen Verlauf nehmen. Bei ausgeprägten Änderungen des Abstandes und insgesamt nicht-parallelem Verlauf ist davon auszugehen, dass sich der Einfluss der Vorkenntnisse in Abhängigkeit von Sprachkenntnissen verändert.

Falls die Doppelte Schwellenhypothese auf die eingesetzten Tests zutrifft, müsste sich folgendes Muster ergeben: Der Abstand der Kurven ist am unteren und am oberen Ende

gering. Im mittleren Bereich ist der Abstand größer. Dann wäre der Einfluss der Vorkenntnisse auf die Leistungen im Leseverstehen mit Fachbezug für Kandidaten mit geringen und hohen Deutschkenntnissen gering, für Kandidaten mit mittleren Deutschkenntnissen jedoch ausgeprägt.

Bei der Analyse der Streudiagramme wurde die Beschreibung der Ergebnisse und die Diskussion nicht getrennt, da eine Beschreibung ohne Einordnung der Phänomene nicht sinnvoll ist.

Streudiagramme mit INFLATION und C-TEST

Im Folgenden werte ich die drei Streudiagramme mit INFLATION nach C-TEST aus (siehe Abbildung 37 Abbildung 38 Abbildung 39, Seite 293, 294 und 295). In allen drei Streudiagrammen nehmen die Kurven einen steigenden Verlauf, was den Einfluss der Deutschkenntnisse (nach C-TEST) auf INFLATION verdeutlicht (siehe auch Kapitel 6.2.1, Seite 268). In allen drei Streudiagrammen verlaufen die Kurven nicht parallel, was darauf hindeutet, dass es eine Interaktion zwischen Vorkenntnissen und Deutschkenntnissen gibt.

Im Streudiagramm nach Kurszugehörigkeit/Studienziel (KURS) ist der Einfluss der Vorkenntnisse bei C-TEST-Ergebnissen zwischen 45 und 70 besonders ausgeprägt (Abbildung 37, Seite 293). Im oberen Bereich (C-TEST höher als 70 Prozent) verlaufen die Kurven recht parallel, im unteren Bereich laufen sie aufeinander zu. Dabei ist zu beachten, dass nur zwei Kandidaten in C-TEST weniger als 10 Prozent erzielten. Vor allem aus dem W-Kurs sind im niedrigen Bereich kaum Ergebnisse zu verzeichnen, so dass der Verlauf der gestrichelten Linie von wenigen Ergebnissen stark beeinflusst wird. Dennoch: Wenn sich das Muster aus diesem Streudiagramm in anderen Konstellationen wiederholen würde, könnte man daraus mit einigem Recht eine Bestätigung auf die Schwellenhypothese ableiten, denn das erwartete Muster (größter Abstand im Bereich der mittleren C-TEST Ergebnisse) trifft mehr oder weniger zu.

In den Streudiagrammen, in denen die Ergebnisse nach BEGRIFFE bzw. BEKANNT dargestellt sind, ähnelt sich der Verlauf der Kurven (Abbildung 38, Abbildung 39, Seite 294 und 295). Im Bereich der niedrigen C-TEST-Ergebnisse ist der Abstand groß und

die Steigung eher gering. Das deutet auf einen großen Einfluss der Vorkenntnisse und einen geringen Einfluss der Deutschkenntnisse hin. Allerdings ist die geringe Zahl der Ergebnisse zu berücksichtigen. Bei einem C-TEST-Ergebnis von 40 Prozent steigt die gestrichelte Kurve (Tendenz für Kandidaten mit Vorkenntnissen) stärker an. Die durchgezogene Kurve schwankt im Bereich für mittlere C-TEST-Ergebnisse, steigt dann aber auch mehr oder weniger parallel zur oberen Kurve an. Im Streudiagramm nach BEKANNT ist im oberen Bereich eine gewisse Annäherung der Kurven auszumachen.

Trotz des uneinheitlichen Verlaufs im Bereich der mittleren Ergebnisse in C-TEST interpretiere ich diese beiden Streudiagramme nicht als Bestätigung für die Schwellenhypothese. Die Kurven verlaufen in großen Abschnitten recht parallel, was auf einen konstanten Einfluss der Vorkenntnisse schließen lässt. Interessant ist ein anderer Aspekt: die geringe Steigung der Kurven im Bereich der niedrigen Ergebnisse in C-TEST und die große Steigung im oberen Bereich. Das ist ein Hinweis darauf, dass sich die Ergebnisse im Leseverstehenstest mit Fachbezug vor allem bei fortgeschrittenen Deutschkenntnissen rasch verbessern.

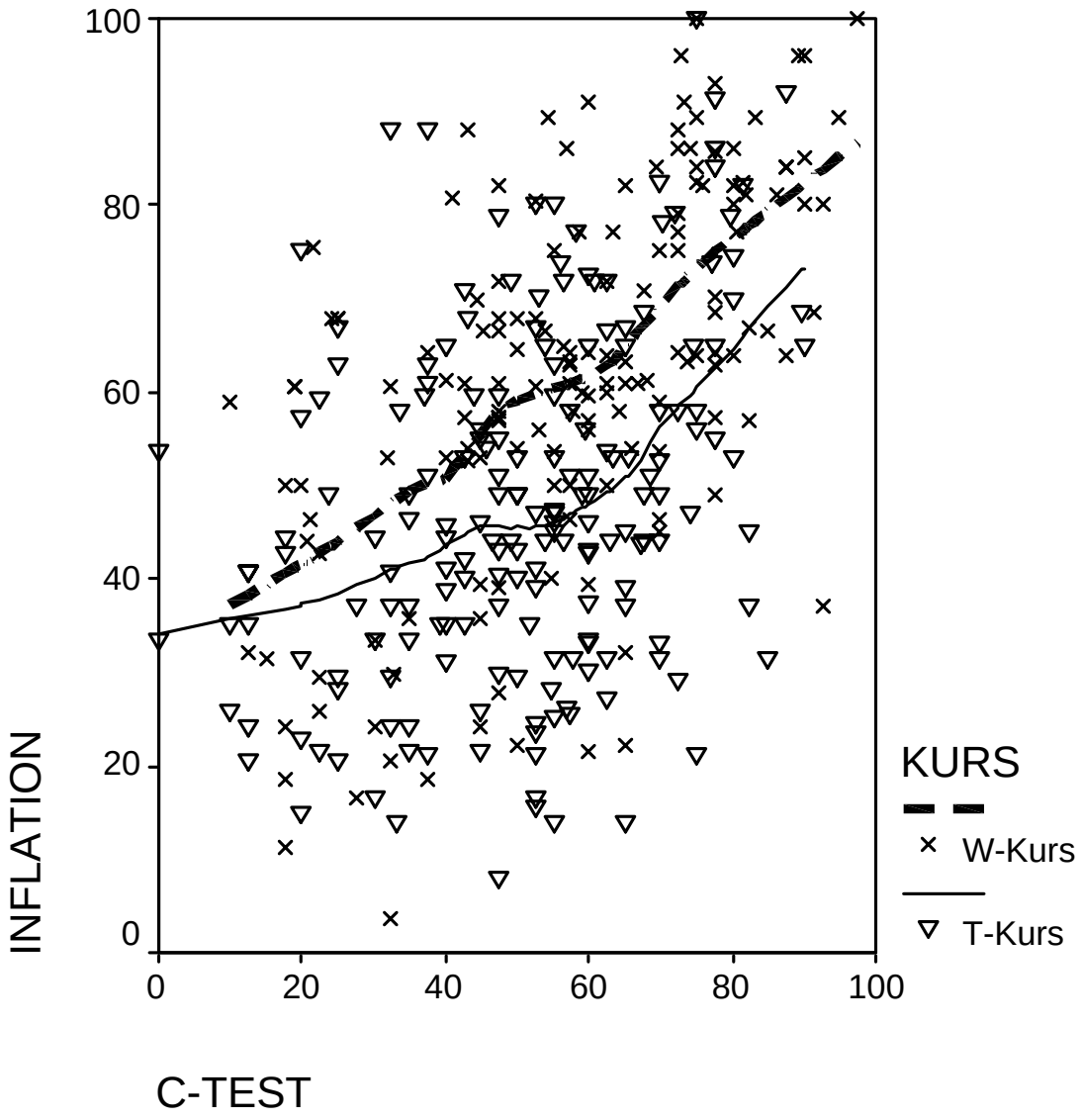


Abbildung 37: Ergebnisse in INFLATION und C-TEST nach KURS (Streudiagramm mit Lowess-Regressionslinien; n = 352)

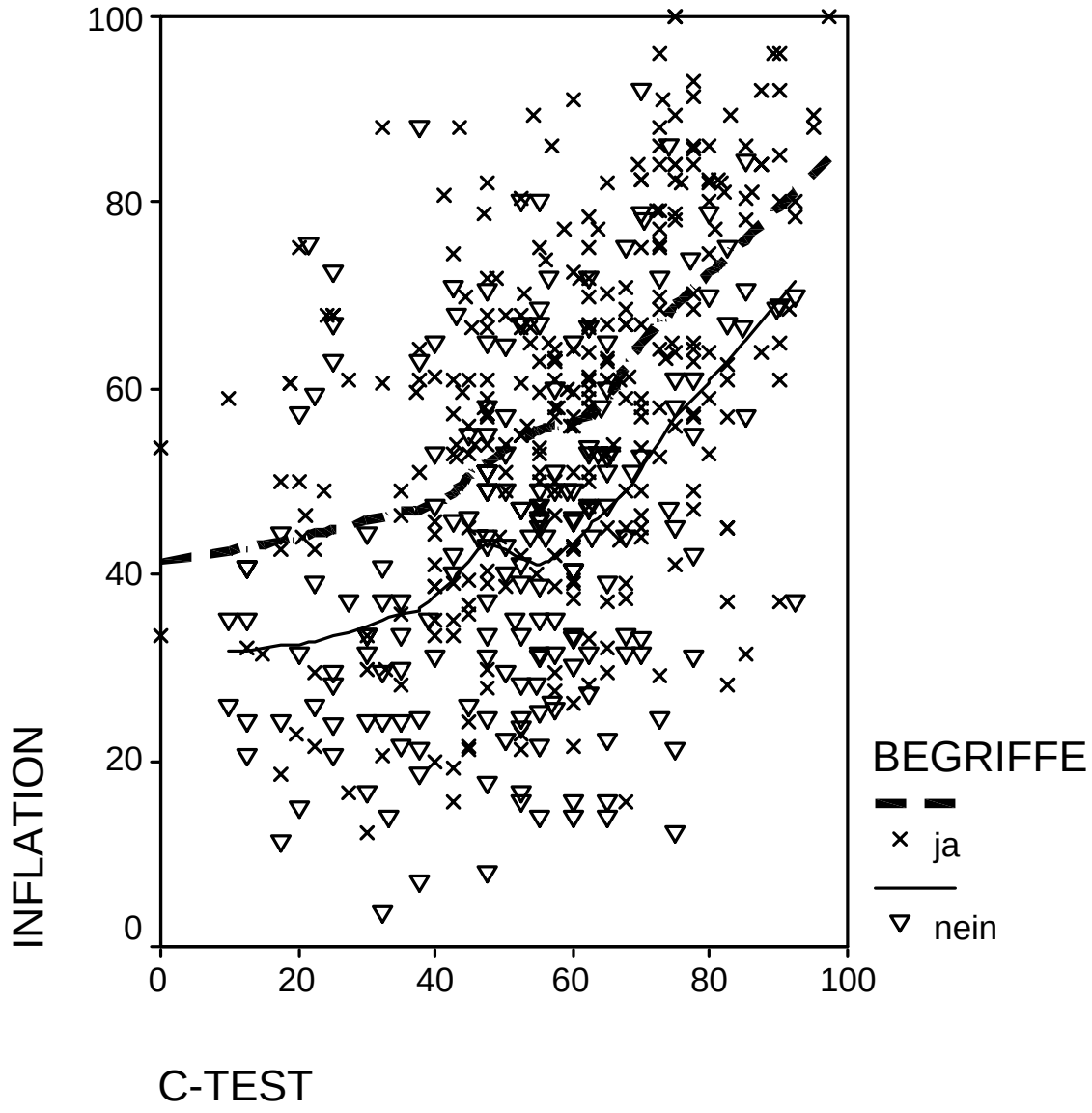


Abbildung 38: Ergebnisse in INFLATION und C-TEST nach BEGRIFFE (Streudiagramm mit Lowess-Regressionslinien; n = 505)

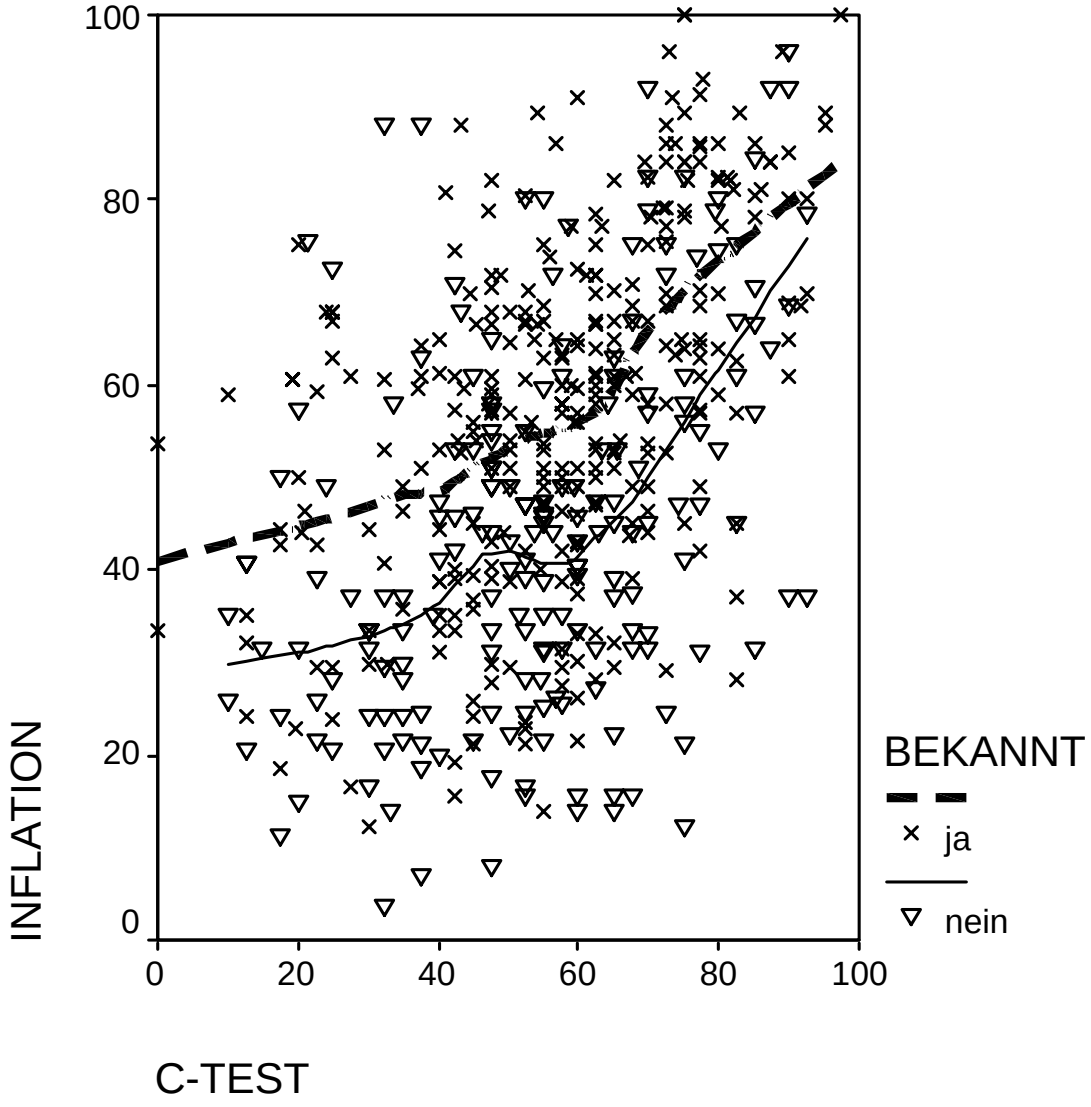


Abbildung 39: Ergebnisse in INFLATION und C-TEST nach BEKANNT (Streudiagramm mit Lowess-Regressionslinien; n = 510)

Streudiagramme mit GESCHWINDIGKEIT und C-TEST

Ich komme nun zu den Streudiagrammen mit den Ergebnissen aus dem Leseverstehenstest "Geschwindigkeit" und C-TEST (Abbildung 40, Abbildung 41, Abbildung 42, Seite 298, 299 und 300).

Im Streudiagramm mit GESCHWINDIGKEIT und KURS verlaufen die Kurven im Bereich der unteren C-TEST-Ergebnisse fast parallel. Bei Ergebnissen in C-TEST zwischen 50 und 75 Prozent nähern sich die Kurven an und verlaufen mit sehr geringem Abstand parallel. Im obersten Leistungsspektrum (C-TEST) gehen die Kurven plötzlich auseinander. Der Einfluss der Kurszugehörigkeit bzw. des Studienziels schwankt über verschiedene Niveaus der Deutschkenntnisse, der Verlauf lässt sich jedoch kaum mit der Schwellenhypothese in Einklang bringen. Im Bereich der unteren Deutschkenntnisse ist der Einfluss der Vorkenntnisse (nach KURS) konstant, im mittleren Bereich nimmt er ab und nimmt im obersten Bereich wieder zu. Selbst wenn man die plötzliche Öffnung der beiden Kurven bei Ergebnissen in C-TEST von über 80 Prozent wegen der geringen Anzahl der Ergebnisse in diesem Bereich nicht in die Bewertung einbezieht, ist die Annäherung bei C-TEST-Ergebnissen zwischen 50 und 70 Prozent kaum schlüssig.

Der Verlauf der Lowess-Regressionslinien im Streudiagramm mit GESCHWINDIGKEIT und C-TEST nach der Kenntnis der Schlüsselbegriffe "Radar" und "Laser" (BEGRIFFE) könnte schon eher als Hinweis auf die Schwellenhypothese interpretiert werden. Im Bereich der C-TEST-Ergebnisse zwischen 35 und 55 Prozent ist der Abstand der Kurven und damit der Einfluss der Vorkenntnisse (nach BEGRIFFE) am größten. Kleiner ist er bei Kandidaten mit sehr fortgeschrittenen Deutschkenntnissen (nach C-TEST), auch bei niedrigen Ergebnissen scheint er sich zu verringern. Allerdings gibt es kaum Kandidaten mit niedrigem Ergebnis in C-TEST, die mit den Schlüsselbegriffen aus dem Text "Geschwindigkeitsmessung" vertraut waren.

Ich komme nun zu dem Streudiagramm, in dem die Ergebnisse nach der Variablen BEKANNT dargestellt werden. Für Kandidaten mit niedrigen und mittleren Deutschkenntnissen (bis zu einem Ergebnis im C-TEST von 65 Prozent) verlaufen beide Kurven parallel. Im oberen Bereich nähern die Kurven sich an. Daraus geht hervor, dass der Einfluss der Vorkenntnisse (nach BEKANNT) für etwa zwei Drittel der Kandidaten

konstant (groß) ist und nur für Kandidaten mit fortgeschrittenen Deutschkenntnissen (C-TEST-Ergebnis über 70 Prozent) leicht abnimmt.

Im Falle des Leseverstehenstests "Geschwindigkeitsmessung" zeigt die mittlere Tendenz der Kurven kein einheitliches Muster. Es gibt Wechselwirkungen zwischen den Vorkenntnissen und dem Niveau der Deutschkenntnisse, der Verlauf dieser Abhängigkeiten hängt jedoch stark von der Methode ab, mit der die Vorkenntnisse erhoben wurden. Das unterschiedliche Verhalten der Lowess-Regressionslinien, also der unterschiedliche Verlauf der mittleren Tendenz ist ein Hinweis auf die eingeschränkte Reliabilität der Variablen zu den Vorkenntnissen. Hätte man Vorkenntnisse nur mit einer Methode erhoben, wäre es kaum möglich, allgemeine Schlussfolgerungen zu ziehen.

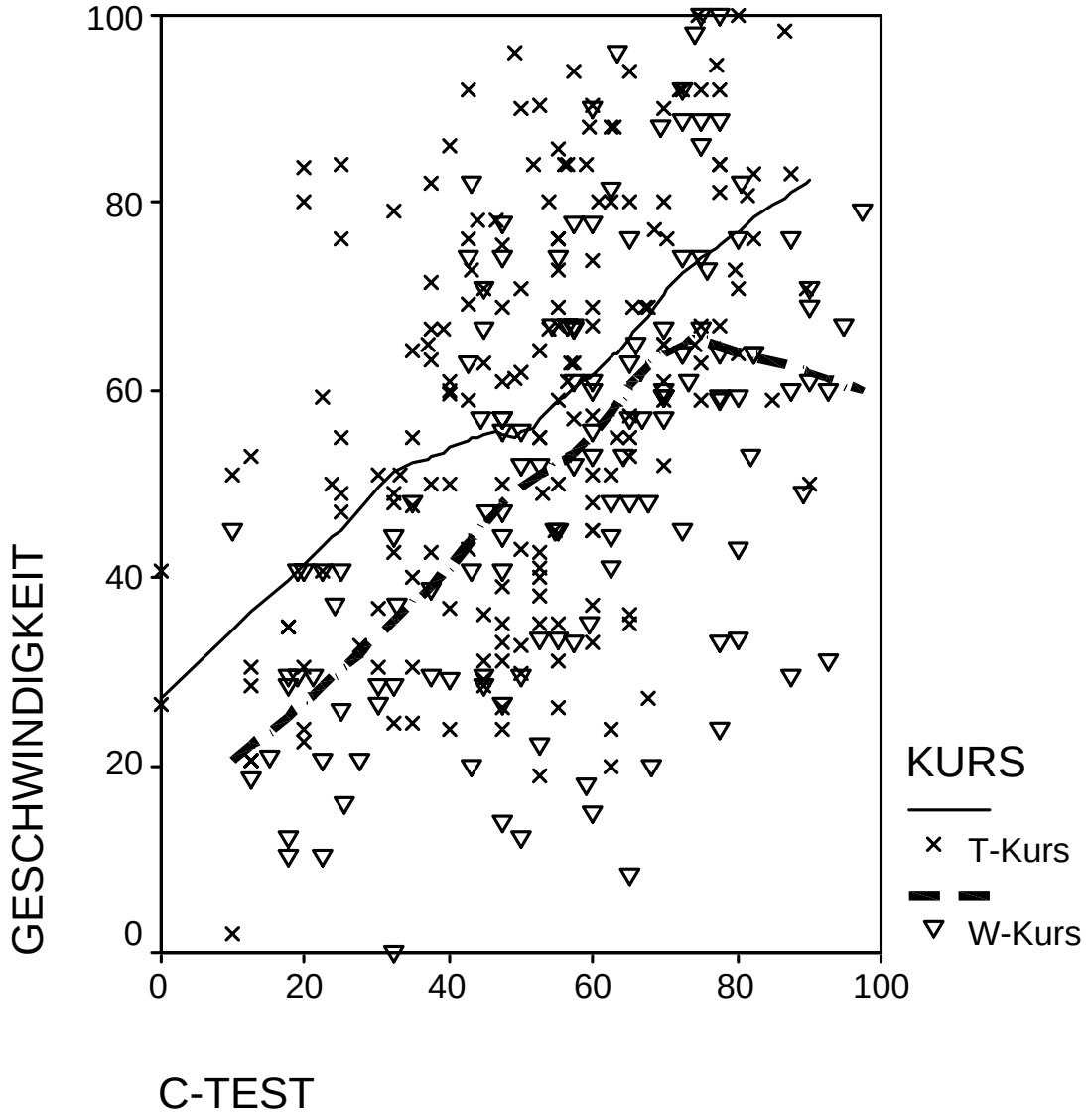


Abbildung 40: Ergebnisse in GESCHWINDIGKEIT und C-TEST nach KURS (Streudiagramm mit Lowess-Regressionslinien; n = 333)

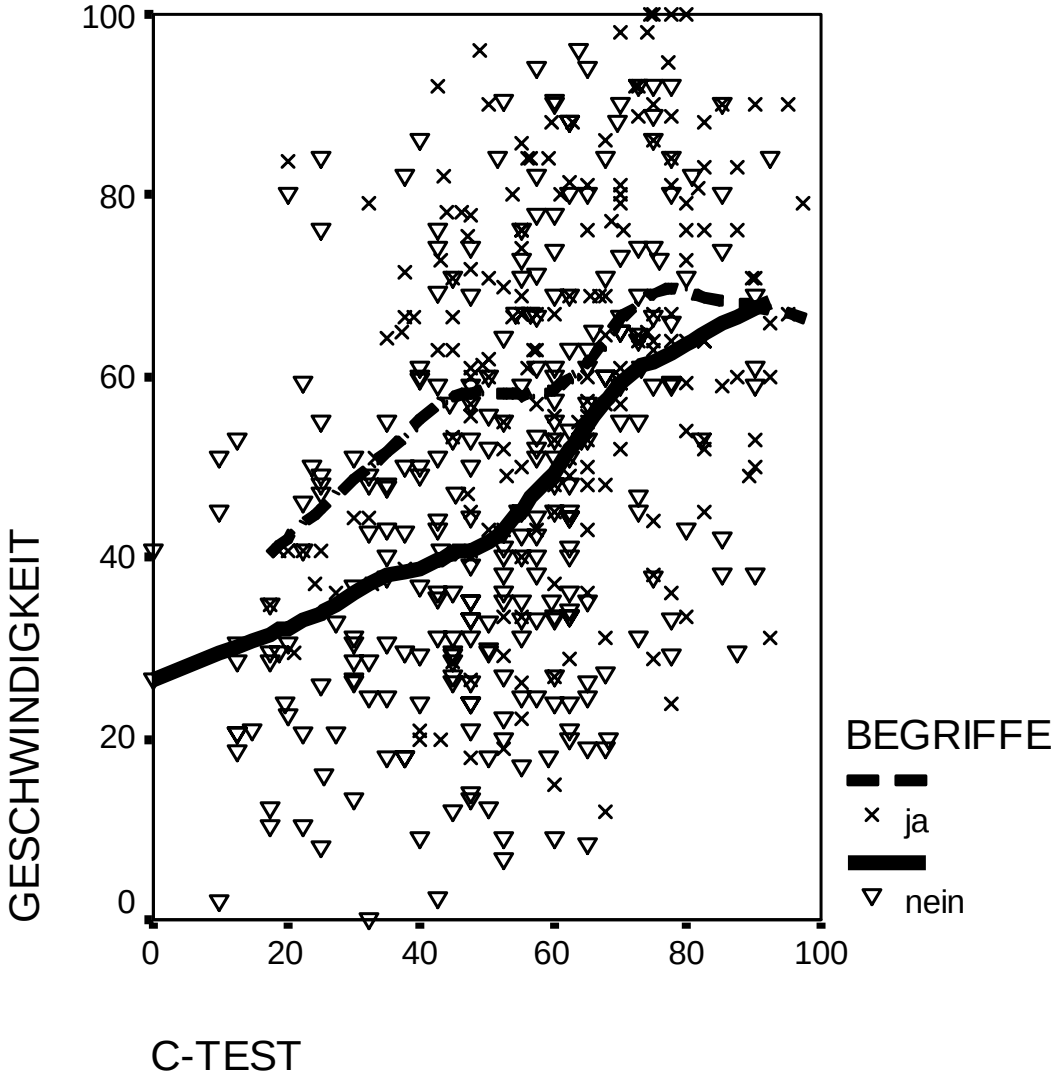


Abbildung 41: Ergebnisse in GESCHWINDIGKEIT und C-TEST nach BEGRIFFE (Streudiagramm mit Lowess-Regressionslinien; n = 486)

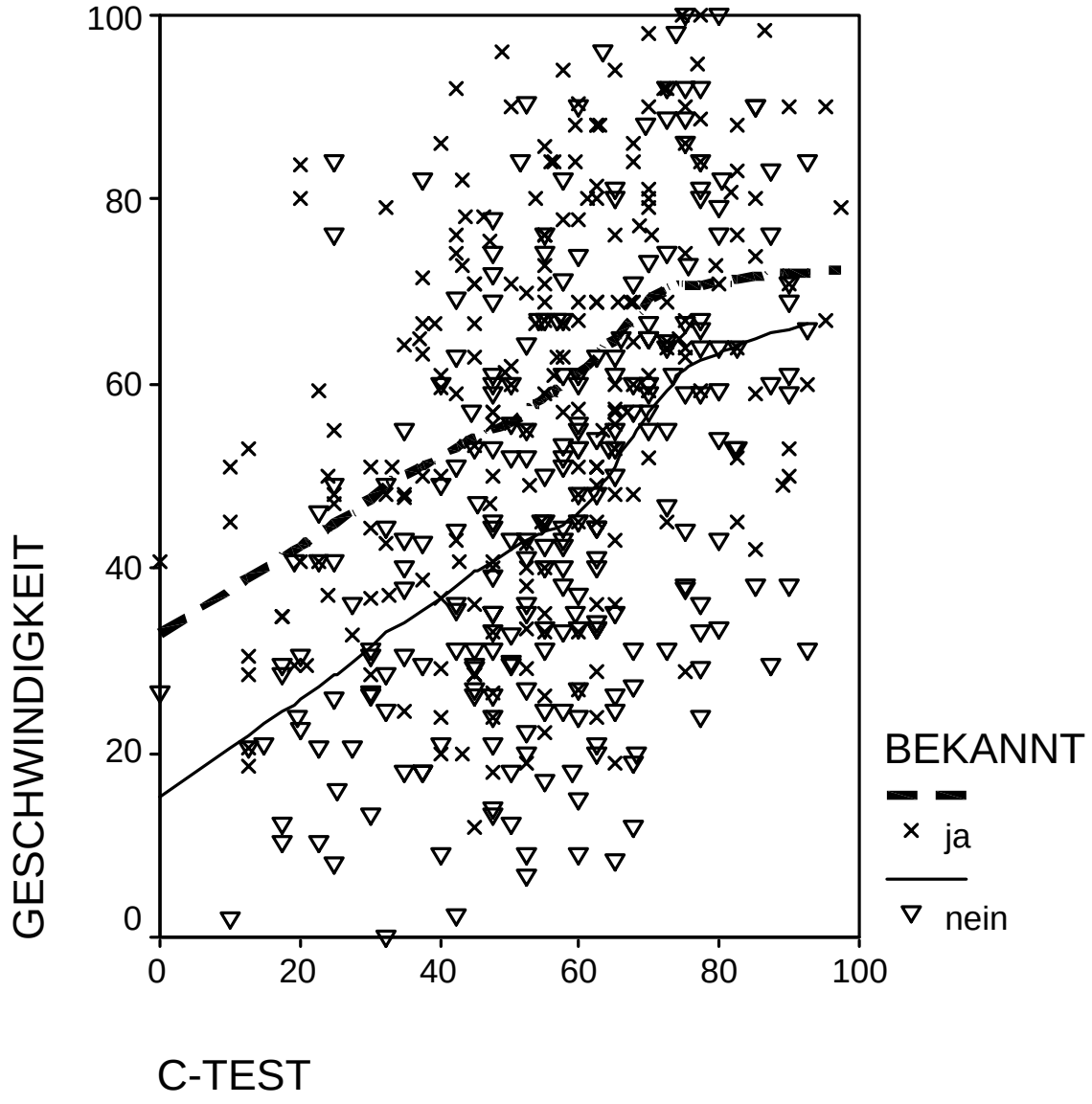


Abbildung 42: Ergebnisse in GESCHWINDIGKEIT und C-TEST nach BEKANNT (Streudiagramm mit Lowess-Regressionslinien; n = 489)

Varianzanalysen mit INFLATION

Während die Streudiagramme vor allem dazu dienen, einen visuellen Eindruck über die Wechselwirkungen zwischen den Vorkenntnisse und den Fremdsprachkenntnissen auf Leistungen in Sprachtests mit Fachbezug zu erhalten, kann die Signifikanz der Wechselwirkungen mit Varianzanalysen ermittelt werden. Für die Berechnungen wurden Kandidaten nach dem Niveau der Deutschkenntnisse, die mit C-TEST auf einer Prozentskala gemessen wurden, in drei Gruppen eingeteilt, in Gruppen mit niedrigen, mittleren und hohen Ergebnissen ("C-Test hoch", "C-Test mittel", "C-Test niedrig"). Die Vorkenntnisse werden mit den in Kapitel 6.1.4 (Seite 257) beschriebenen Variablen Studienziel/Kurszugehörigkeit (KURS), Kenntnis der Schlüsselbegriffe vor dem Lesen (BEGRIFFE) sowie Vorkenntnisse laut Selbstauskunft nach dem Lesen (BEKANNT) eingegeben. Zu den beiden Texten "Inflation" und "Geschwindigkeit" wurden drei Varianzanalysen mit den jeweiligen Variablen zu den Vorkenntnissen durchgeführt. Bei welchen Niveaustufen verändert sich gegebenenfalls die Rolle der Vorkenntnisse? Die Einteilung der Kandidaten in drei Gruppen nach dem Niveau der Deutschkenntnisse geschah nicht nach einem festen Schema. Durch unterschiedliche Einteilungen sollte das Niveau deutlich werden, ab dem sich die Rolle der Vorkenntnisse ändert.

Wie auch die anderen Tabellen in diesem Abschnitt zeigt die Tabelle 54 die Ergebnisse in folgender Form: Betrachtet wurden neben der Grundgesamtheit ("alle") einzelne Teilgruppen (z. B. Ergebnis "unter 50 %" in C-TEST). So wurden beispielsweise nur die Ergebnisse der Kandidaten betrachtet, welche im C-Test weniger als 50 Prozent erzielten. Das sind 136 Kandidaten, von denen 84 den T-Kurs besuchen und 52 den W-Kurs. Die Tabelle zeigt weiter die Mittelwerte dieser Teilgruppen in einem der beiden Leseverstehenstests. Mit einer einfaktoriellen Varianzanalyse wurde ermittelt, ob die Unterschiede zwischen den Mittelwerten signifikant sind (*t*-Tests zu unabhängigen Stichproben führen zu vergleichbaren Ergebnissen). Ein nicht signifikanter Mittelwertunterschied deutet darauf hin, dass ein Einfluss der Vorkenntnisse auf die Leistungen im Leseverstehenstest mit geringem Fachlichkeitsgrad nicht nachgewiesen werden kann. Wenn die Varianzanalyse oder der *t*-Test die Unterschiede zwischen den Mittelwerten als signifikant ausweist, ist davon auszugehen, dass die Leistungen im Leseverstehenstest von den Vorkenntnissen beeinflusst werden. Wenn die Doppelte Schwellen-

hypothese zutreffen sollte, dürfte der Unterschied der Mittelwerte von Gruppen mit niedrigen und hohen Deutschkenntnissen nicht signifikant sein.

Tabelle 54: Ergebnisse im Leseverstehenstest "Inflation" nach Deutschkenntnissen und Vorkenntnissen (Variable KURS)

Sprach-kompetenz (C-TEST)	Vorkenntnisse (KURS)	Anzahl (n)	Mittelwerte (*AM*) in %	Differenz in % Signifikanz (*p*)
Gruppen mit niedrigen Deutschkenntnissen (nach C-TEST)				
< 40 %	Nein (T-Kurs)	49	40 %	-5 %
	Ja (W-Kurs)	26	35 %	F = 1.024; nicht sign.
< 50 %	Nein (T-Kurs)	84	42 %	5 %
	Ja (W-Kurs)	52	47 %	F = 1.934; nicht sign.
< 55 %	Nein (T-Kurs)	105	42 %	7 %
	Ja (W-Kurs)	63	49 %	F = 5.626; $p < .05$
< 60 %	Nein (T-Kurs)	128	43 %	9 %
	Ja (W-Kurs)	77	52 %	F = 9.779; $p < .01$
Gruppen mit fortgeschrittenen Deutschkenntnissen (nach C-TEST)				
> 60 %	Nein (T-Kurs)	55	57 %	15 %
	Ja (W-Kurs)	71	72 %	F = 23.300; $p < .01$
> 70 %	Nein (T-Kurs)	27	65 %	13 %
	Ja (W-Kurs)	49	78 %	F = 11.412; $p = .01$
> 75 %	Nein (T-Kurs)	17	68 %	8 %
	Ja (W-Kurs)	34	76 %	F = 3.108; nicht sign.

Tabelle 55: Ergebnisse im Leseverstehenstest "Inflation" nach Deutschkenntnissen und Vorkenntnissen (Variable BEGRIFFE)

Sprach-kompetenz (C-TEST)	Schlüsselbegriffe bekannt? BEGRIFFE	Anzahl (n)	Mittelwerte (*AM*) in %	Differenz in % Signifikanz (*p*)
alle	ja/nein	307/198	58 % - 44 %	14 % F = 68,457; $p < 0,01$
Gruppen mit niedrigen Deutschkenntnissen (nach C-TEST)				
unter 30 %	ja/nein	18/26	40 % - 37 %	3 % F = 0,438; nicht signifikant
unter 40 %	ja/nein	32/47	44 % - 35 %	9 % F = 4,667; $p < 0,05$
unter 50 %	ja/nein	80/75	49 % - 39 %	10 % F = 12,497; $p < 0,01$
Gruppen mit fortgeschrittenen Deutschkenntnissen (nach C-TEST)				
über 60 %	Ja/nein	149/62	66 % - 52 %	14 % F = 26,223; $p < 0,01$
über 70 %	Ja/nein	95/29	74 % - 62 %	12 % F = 10,061; $p < 0,01$
über 80 %	Ja/nein	43/11	74 % - 67 %	7 % F = 1,613; nicht signifikant

Die ersten Varianzanalysen beziehen folgende Variablen ein: Ergebnisse im Leseverstehenstest "Inflation" als abhängige Variable sowie Deutschkenntnisse und Vorkenntnisse nach den bekannten Variablen (KURS, BEGRIFFE, BEKANNT siehe Tabelle 54, Tabelle 55, Tabelle 56). Kandidaten mit einem niedrigen Ergebnis im C-Test erzielen unabhängig von ihrem Studienwunsch (KURS) im Leseverstehenstest "Inflation" Ergebnisse, die auf einem vergleichbaren, niedrigen Niveau liegen. Die Unterschiede im Umfang von zwei bis fünf Prozent zwischen den Mittelwerten sind nicht signifikant. Die Kandidaten aus T-Kursen erzielten sogar etwas höhere Ergebnisse als Kandidaten aus W-Kursen. Die Schwelle, ab der die Kandidaten Vorkenntnisse (nach Studienziel) Gewinn bringend einsetzen können, scheint bei einem Ergebnis von ungefähr 50 Prozent im C-Test zu liegen. Im oberen Bereich sind unterschiedliche Signifikanzniveaus zu beobachten. Bei den Kandidaten mit einem Ergebnis von über 75 Prozent im C-Test ist der Unterschied zwischen den Kursen bzw. zwischen den Studienzielen nicht signifikant. Ob dieses Ergebnis als Beleg dafür gedeutet werden kann, dass Kandidaten mit fortgeschrittenen Deutschkenntnissen beim Leseverstehenstest mit geringem Fachlichkeitsgrad unabhängig von ihren Vorkenntnissen ein hohes Ergebnis erzielen, ist fraglich. Die fehlende Signifikanz könnte auch eine Folge der kleinen Probandenzahl sein. Zur Teilgruppe mit besonders hohen Ergebnissen im C-Test gehören nur 17 Kandidaten aus T-Kursen. Der Unterschied zwischen den mittleren Ergebnissen beträgt acht Prozent. Wenn man Gruppen mit Kandidaten bildet, die ein mittleres Ergebnis im C-Test erzielten, ist jeweils ein signifikanter Unterschied zwischen den Ergebnismittelwerten zu beobachten. Diese Ergebnisse wurden nicht abgebildet.

Wenn man Vorkenntnisse am Studienziel und der Kurszuweisung im Studienkolleg festmacht, spielen Vorkenntnisse beim Leseverstehenstest "Inflation" nur bei Kandidaten eine Rolle, die mindestens über eine mittlere Deutschkompetenz verfügen (Ergebnis im C-Test über 50 Prozent). Auch bei Kandidaten mit weit fortgeschrittenen Deutschkenntnissen scheinen die Vorkenntnisse eine Rolle zu spielen.

Ändern sich die beschriebenen Zusammenhänge, wenn die Vorkenntnisse nicht über das Studienziel, sondern über die Kenntnis der Schlüsselbegriffe vor dem Lesen erhoben werden (Variable BEGRIFFE)? Zunächst muss in Erinnerung gerufen werden, dass die Unterscheidung nach Kenntnis der Schlüsselbegriffe vor dem Lesen zu wesentlich

stärkeren Leistungsunterschieden im Leseverstehen mit geringem Fachlichkeitsgrad führt. Die Mittelwertunterschiede zwischen beiden Gruppen sind daher besonders groß: In beiden Leseverstehenstests erzielen Testteilnehmer, welche mit den Schlüsselbegriffen bereits vertraut waren, ein im Mittel um 15 Prozent höheres Ergebnis (Tabelle 55, Seite 302). Vor diesem Hintergrund ist es nicht überraschend, dass es auch bei den Ergebnissen von Teilgruppen nach dem Niveau der Deutschkenntnisse zu signifikanten Unterschieden zwischen den Leistungen in den Leseverstehenstests kommt.

Beim Leseverstehenstest "Inflation" sind die Mittelwertunterschiede zwischen den Gruppen (BEGRIFFE) allein am ganz oberen und ganz unteren Ende des Spektrums nicht mehr signifikant.

Schließlich analysiere ich, ob eine Wechselwirkung zwischen Vorkenntnissen, Deutschkenntnissen und Leistungen in Leseverstehenstest mit geringem Fachlichkeitsgrad deutlich wird, wenn die Vorkenntnisse laut Selbstauskunft nach dem Lesen erhoben werden (Variable BEKANNT). Beim Leseverstehenstest "Inflation" ist wiederum ein signifikanter Einfluss der Vorkenntnisse zu beobachten. Bei den (zahlenmäßig kleinen) Gruppen mit besonders geringen oder besonders hohen Deutschkenntnissen ist der Unterschied zwischen den Mittelwerten nicht signifikant, obwohl er zehn bzw. sieben Prozent beträgt (Tabelle 56, Seite 305).

Tabelle 56: Ergebnisse im Leseverstehenstest "Inflation" nach Deutschkenntnissen und Vorkenntnissen (Variable BEKANNT)

Sprach-kompetenz (C-TEST)	**Thema bekannt? BEKANNT**	**Anzahl (n)**	**Mittelwerte (*AM*) in %**	**Differenz in % Signifikanz (*p*)**
Alle	Ja/nein	314/196	58 % - 45 %	13 % F = 57,003; $p < 0,01$
Gruppen mit niedrigen Deutschkenntnissen (nach C-TEST)				
unter 30 %	Ja/nein	31/22	49 % - 35 %	14 % F = 7,377; $p < 0,01$
unter 40 %	Ja/nein	46/50	47 % - 34 %	14 % F = 13,320; $p < 0,01$
unter 50 %	Ja/nein	101/83	49 % - 38 %	11 % F = 18,593; $p < 0,01$
Gruppen mit fortgeschrittenen Deutschkenntnissen (nach C-TEST)				
über 60 %	Ja/nein	156/77	65 % - 51 %	14 % F = 26,582; $p < 0,01$
über 70 %	Ja/nein	89/44	72 % - 60 %	12 % F = 13,499; $p < 0,01$
über 80 %	Ja/nein	35/20	75 % - 67 %	8 % F = 2,734; nicht signifikant

Varianzanalysen mit GESCHWINDIGKEIT

Auch mit den Ergebnissen im Leseverstehenstest "Geschwindigkeitsmessung" wurden Varianzanalysen durchgeführt (siehe Tabelle 57, Tabelle 58, Tabelle 59, Seite 307 und 308).

- Wenn die Leistungen der Kandidaten nach der Kurszuweisung bzw. dem Studienziel unterschieden werden (Variable KURS, Tabelle 57), ergibt sich folgendes Muster: Die Ergebnisse der Kandidaten aus den T-Kursen liegen zwar im Mittel stets über denen der Kandidaten aus den W-Kursen, aber bei mittleren Deutschkenntnissen betragen diese Unterschiede nur drei bis sechs Prozent und sind nicht signifikant. Diese Feststellung ist angesichts der relativ großen Probandenzahlen im mittleren Bereich bemerkenswert. Im oberen und unteren Bereich sind die Unterschiede zwischen den Mittelwerten demgegenüber groß und auf dem Niveau 99 Prozent signifikant.

- Bei einer Betrachtung der Ergebnisse nach der Variablen BEGRIFFE (Vertrautheit mit den Schlüsselbegriffen "Radar" und/oder "Laser") liegen die Leistungen der Kandidaten mit Vorkenntnissen stets über denen der anderen Gruppe, jedoch verringert sich die Differenz bei Kandidaten mit besonders hohen Deutschkenntnissen auf unter zehn Prozent (Tabelle 58). Diese Unterschiede sind nicht signifikant, was allerdings auch eine Folge der kleinen Anzahl sein könnte.

- Ein ähnliches Muster ist auch zu beobachten, wenn die Ergebnisse nach der Variablen BEKANNT betrachtet werden (Frage nach dem Lesen: mit dem Thema vertraut?, siehe Tabelle 59). Die Differenz der Mittelwerte zwischen den Gruppen ist nach der Variable BEKANNT am geringsten. Bei Gruppen mit Ergebnissen in C-TEST von über 70 Prozent beträgt der Mittelwertunterschied nur noch 6 Prozent und ist nicht mehr signifikant. Bei allen anderen Gruppen ist der Einfluss der Vorkenntnisse ausgeprägter (10 bis 15 Prozent).

Zumindest im oberen Bereich entsprechen die Ergebnisse nach BEKANNT und BEGRIFFE annähernd der Erwartung der Schwellenhypothese. Am oberen (sprachlichen) Leistungsspektrum (über 70 bzw. 80 Prozent im C-Test) trifft man einen abnehmenden Einfluss der Vorkenntnisse an. Die Mittelwertunterschiede gehen von zehn

Prozent auf sieben bzw. fünf zurück und sind nicht signifikant. Eine Unterscheidung der Gruppen nach KURS führt zu Ergebnissen, die der Doppelten Schwellenhypothese widersprechen.

Tabelle 57: Ergebnisse im Leseverstehenstest "Geschwindigkeit" nach Deutschkenntnissen und Vorkenntnissen (Variable KURS)

Sprach-kompetenz (C-TEST)	**Vorkenntnisse (KURS)**	**Anzahl (n)**	**Mittelwerte (*AM*) in %**	**Differenz in % Signifikanz (*p*)**
Gesamt	Ja (T-Kurs)	192	59 %	7 %
	Nein (W-Kurs)	141	52 %	F = 9,830; $p < .01$
Gruppen mit niedrigen Deutschkenntnissen (nach C-TEST)				
< 30 %	Ja (T-Kurs)	26	42 %	21 %
	Nein (W-Kurs)	13	21 %	F = 11.099; $p < .01$
< 40 %	Ja (T-Kurs)	49	47 %	21 %
	Nein (W-Kurs)	23	26 %	F = 21.188; $p < .01$
< 50 %	Ja (T-Kurs)	84	50 %	13 %
	Nein (W-Kurs)	46	37 %	F = 12.185; $p < .01$
Gruppen mit mittleren Deutschkenntnissen (nach C-TEST)				
47 – 65 %	Ja (T-Kurs)	73	58 %	6 %
	Nein (W-Kurs)	44	52 %	F = 2.478; nicht sign.
40 – 75 %	Ja (T-Kurs)	113	60 %	5 %
	Nein (W-Kurs)	78	55 %	F = 3.193; nicht sign.
50 – 75 %	Ja (T-Kurs)	79	62 %	5 %
	Nein (W-Kurs)	53	57 %	F = 1.803; nicht sign.
Gruppen mit fortgeschrittenen Deutschkenntnissen (nach C-TEST)				
> 60 %	Ja (T-Kurs)	55	70 %	8 %
	Nein (W-Kurs)	62	62 %	F = 5.121; $p < .05$
> 70 %	Ja (T-Kurs)	27	78 %	14 %
	Nein (W-Kurs)	41	64 %	F = 9.183; $p < .01$
> 80 %	Ja (T-Kurs)	8	75 %	14 %
	Nein (W-Kurs)	15	61 %	F = 4.430; $p < .05$

Tabelle 58: Ergebnisse im Leseverstehenstest "Geschwindigkeit" nach Deutschkenntnissen und Vorkenntnissen (Variable BEGRIFFE)

Sprach-kompetenz (C-TEST)	**Schlüsselbegriffe bekannt? BEGRIFFE**	**Anzahl (n)**	**Mittelwerte (*AM*) in %**	**Differenz in % Signifikanz (*p*)**
alle	ja/nein	199/287	61 % - 47 %	14 % F = 53,180; $p < 0,01$
Gruppen mit niedrigen Deutschkenntnissen (nach C-TEST)				
unter 40 %	ja/nein	14/63	56 % - 36 %	20 % F = 13,662; $p < 0,01$
unter 50 %	ja/nein	40/111	59 % - 39 %	20 % F = 31,734; $p < 0,01$
Gruppen mit fortgeschrittenen Deutschkenntnissen (nach C-TEST)				
über 60 %	ja/nein	94/84	68 % - 58 %	10 % F = 8,597; $p < 0,01$
über 70 %	ja/nein	59/39	70 % - 63 %	7 % F = 3,419; nicht signifikant
über 80 %	ja/nein	26/13	67 % - 62 %	5 % F = 0,978; nicht signifikant

Tabelle 59: Ergebnisse im Leseverstehenstest "Geschwindigkeit" nach Deutschkenntnissen und Vorkenntnissen (Variable BEKANNT)

Sprach-kompetenz (C-TEST)	**Thema bekannt? BEKANNT**	**Anzahl (n)**	**Mittelwerte (*AM*) in %**	**Differenz in % Signifikanz (*p*)**
Alle	ja/nein	230/259	59 % - 47 %	12 % F = 39,974; $p < 0,01$
Gruppen mit niedrigen Deutschkenntnissen (nach C-TEST)				
unter 30 %	ja/nein	23/29	43 % - 28 %	15 % F = 9,390; $p < 0,01$
unter 40 %	ja/nein	43/52	47 % - 30 %	17 % F = 22,112; $p < 0,01$
unter 50 %	ja/nein	86/94	50 % - 35 %	15 % F = 25,514; $p < 0,01$
Gruppen mit fortgeschrittenen Deutschkenntnissen (nach C-TEST)				
über 60 %	ja/nein	103/119	66 % - 56 %	10 % F = 10,508; $p < 0,01$
über 70 %	ja/nein	55/67	70 % - 64 %	6 % F = 3,059; nicht signifikant
über 80 %	ja/nein	23/27	67 % - 63 %	4 % F = 0;400; nicht signifikant

Streudiagramm: Differenz der Ergebnisse (nach KURS)

Als letzte Analyse zu Leseverstehenstests mit Fachbezug möchte ich noch ein Streudiagramm mit der Differenz der Ergebnisse aus INFLATION und GESCHWINDIGKEIT vorstellen (Abbildung 43, Seite 310). Die Ergebnisse in C-TEST sind entlang der X-Achse dargestellt. Die Y-Achse zeigt die Differenz der Ergebnisse aus INFLATION und GESCHWINDIGKEIT. Für einen Kandidaten, der 80 Prozent in INFLATION und 50 Prozent in GESCHWINDIGKEIT erzielte, beträgt die Differenz 30 Prozent. Wenn die Differenz positiv ist, waren die Ergebnisse in INFLATION höher, bei einer negativen Differenz wurde in GESCHWINDIGKEIT ein höheres Ergebnis erzielt.

Ein derartiges Diagramm lässt sich nur mit der Variablen KURS erstellen, da das Merkmal Kurszugehörigkeit bzw. das Studienziel die Kandidaten in zwei feste Gruppen teilt. Die anderen beiden Variablen unterscheiden sich jeweils mit Blick auf INFLATION oder GESCHWINDIGKEIT. Es ist also möglich, dass Kandidaten nach BEGRIFF oder BEKANNT über Vorkenntnisse zu beiden oder zu keinem der beiden Texte verfügt.

Im Streudiagramm sind die Ergebnisse der Kandidaten außerdem nach der Variablen KURS unterschiedlich dargestellt: Ergebnisse von Kandidaten aus dem T-Kurs mit dem Studienziel technische Studiengänge werden durch Dreiecke dargestellt, Ergebnisse von Kandidaten mit dem Studienziel Wirtschaft durch Kreuzchen. Es ist zu erkennen, dass sich mehr Dreiecke im unteren Bereich des Diagramms befinden und mehr Kreuzchen im oberen Bereich. Viele Kandidaten aus T-Kursen erzielten ein besseres Ergebnis in GESCHWINDIGKEIT und Kandidaten aus W-Kursen in INFLATION.

Auch dieses Streudiagramm enthält wieder Lowess-Regressionslinien. Der Abstand zwischen den Kurven veranschaulicht wiederum den Unterschied der Kandidaten mit unterschiedlichen Studienzielen. Wenn es keine Wechselwirkung zwischen Vorkenntnissen (nach KURS) und Deutschkenntnissen (nach C-TEST) gäbe, müssten beide Kurven gerade verlaufen.

Wenn die Ergebnisse die Doppelte Schwellenhypothese unterstützten, müssten die Kurven am oberen und unteren Ende gegen Null gehen, d. h. der Einfluss der Vorkenntnisse für Kandidaten mit geringen und hohen Deutschkenntnissen nachlassen. Im mittleren Bereich müsste der Abstand vom Nullpunkt groß sein. Dieser Zusammenhang

trifft auf eine der beiden Kurven zu, nämlich auf die Kurve, welche die zentrale Tendenz der Ergebnisse der Kandidaten aus T-Kursen abbildet. Diese Kurve zeigt, dass die Kurszugehörigkeit für Kandidaten mit besonders hohen oder besonders niedrigen Deutschkenntnissen keine Rolle spielt. Für Kandidaten mit mittleren Sprachkenntnissen ist der Effekt ausgeprägt – wie von der Doppelten Schwellenhypothese angenommen.

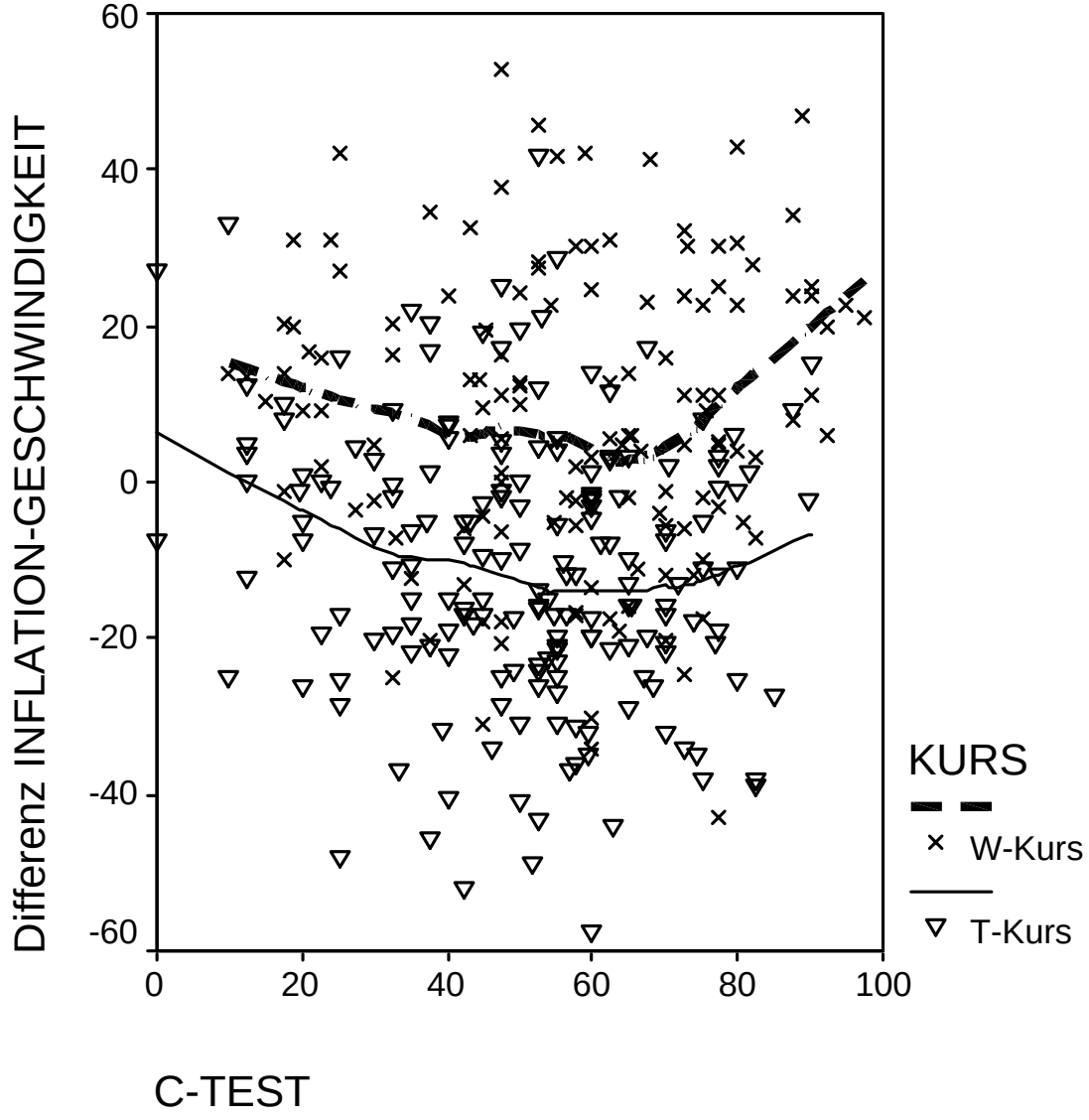


Abbildung 43: Differenz der Ergebnisse in INFLATION und GESCHWINDIGKEIT nach C-TEST (Streudiagramm mit Markierung nach KURS und mit LOWESS Regressionslinien)

Der gegenteilige Effekt ist allerdings bei der Kurve für die Ergebnisse von Kandidaten aus W-Kursen zu beobachten: Kandidaten mit mittleren Ergebnissen in C-TEST (40 bis 70 Prozent) erzielen in beiden Tests ähnliche Ergebnisse, während Kandidaten mit niedrigen und besonders hohen Deutschkenntnissen im Test mit einem Thema aus ihrem zukünftigen Studienfach, "Inflation" ein besseres Ergebnis erzielen als im Test mit einem "fremden" Thema. Dies ist ein Widerspruch zur Doppelten Schwellenhypothese.

Diskussion: Doppelte Schwellenhypothese

Nach den Erwartungen der "Doppelten Schwellenhypothese" hätte eine Wechselwirkung der Vorkenntnisse in Abhängigkeit der Deutschkenntnisse auftreten müssen: Der Einfluss der Vorkenntnisse bei Kandidaten mit mittleren Deutschkenntnissen hätte hoch, bei Kandidaten mit geringen und weit fortgeschrittenen Deutschkenntnissen jedoch deutlich geringer ausgeprägt sein müssen.

Betrachtet man den Einfluss der Vorkenntnisse in Abhängigkeit von den Deutschkenntnissen, so waren bei den eingesetzten Leseverstehenstests unterschiedliche Effekte zu beobachten. Beim Leseverstehenstest "Inflation" war eine ausgeprägte Wechselwirkung mit Vorkenntnissen (nach KURS) zu beobachten, welche dem Muster der Doppelten Schwellenhypothese ähnelt. Allerdings ist auch bei besonders fortgeschrittenen Deutschsprechern ein Einfluss der Vorkenntnisse signifikant. Bei der zahlenmäßig größten Gruppe, den Kandidaten mit einem mittleren Ergebnis im C-Test, ist ein signifikanter Einfluss der Vorkenntnisse auf die Leistungen in den Leseverstehenstests zu beobachten. Dies gilt im Prinzip unabhängig von der Erhebungsmethode der Vorkenntnisse und unabhängig vom jeweiligen Leseverstehenstest.

Nur bei wenigen Kandidaten am unteren Ende des Leistungsspektrums (Deutschkenntnisse nach C-TEST) gibt es Grund zur Annahme, dass der Einfluss der Deutschkenntnisse auf die Leistungen in Leseverstehenstests mit Fachbezug nachlässt bzw. nicht mehr nachzuweisen ist. Im oberen Leistungsspektrum ist das Ergebnis noch weniger eindeutig. Auch Kandidaten mit weit fortgeschrittenen Deutschkenntnissen dürften von ihren Vorkenntnissen Gebrauch machen. Nur in wenigen Fällen dürfte es dazu kommen sein, dass die Vorkenntnisse keine Rolle für das Ergebnis im Lese-

verstehenstest mit geringem Fachlichkeitsgrad spielten. Es gibt einzelne Ergebnisse der Varianzanalysen, welche darauf hindeuten, dass es auch eine obere Schwelle gibt, ab der Vorkenntnisse nicht mehr eingesetzt werden. Doch dies könnte auch eine Folge der geringen Anzahl sein.

Beim Leseverstehenstest "Geschwindigkeit" spielten die Vorkenntnisse auch bei Kandidaten mit geringen Deutschkenntnissen eine große Rolle. Diese ist je nach Erhebungsmethode bisweilen sogar größer als bei Kandidaten mit mittleren Deutschkenntnissen. Auffällig war schließlich, dass die Kurszuordnung bzw. das Studienziel bei Kandidaten (Variable KURS) mit *mittleren* Deutschkenntnissen keinen signifikanten Einfluss auf das Ergebnis im Leseverstehenstest "Geschwindigkeit" hatte. Es war im Gegenteil so, dass die Kurszuordnung bzw. das Studienziel bei den Kandidaten mit besonders geringen oder besonders hohen Deutschkenntnissen einen Einfluss auf das Ergebnis im Leseverstehenstest haben. Diese Auffälligkeit mag jedoch der Besonderheit der Variablen KURS zuzuschreiben sein (siehe Kapitel 6.1.4). Festzuhalten ist, dass die Ergebnisse im Leseverstehenstest "Geschwindigkeit" nicht dem Muster folgten, welche nach der "Doppelten Schwellenhypothese" zu erwarten waren.

6.3. Zusammenfassung und Ausblick

Übersicht: Kapitel 6.3
In diesem Kapitel stelle ich meine Forschungsergebnisse zu den Fragen dar, die zu Beginn des Kapitels 6.1 formuliert wurden. Es folgen Überlegungen zu möglichen Ursachen und Auswirkungen. Das Kapitel schließt mit einem Exkurs über die Auswirkungen sprachwissenschaftlicher Forschung. Anlass ist die Diskussion um eine der Schwellenhypothese vergleichbare Theorie, Cummins' Interdependenzhypothese, mit der bilinguale Erziehungsprogramme begründet werden.

Fragestellung 1
Erzielen ausländische Studienbewerber bessere Ergebnisse in Leseverstehenstests mit Fachbezug, wenn sie über Vorkenntnisse verfügen?

Zum Einfluss der Vorkenntnisse wurden folgende Beobachtungen gemacht:

Vorkenntnisse hatten einen signifikanten Einfluss auf die Leistungen im Leseverstehenstests mit Fachbezug. Die Ergebnisse bestätigen, dass Vorkenntnisse eine wichtige Rolle für die Ergebnisse in Leseverstehenstests mit Fachbezug spielen. Unabhängig von der Erhebungsmethode wurde folgender Trend bestätigt: Kandidaten, welche (laut Studienziel/Kurszuweisung oder laut Kenntnis der Schlüsselbegriffe oder laut Selbstauskunft nach dem Lesen) über Vorkenntnisse verfügen (müssten), erzielten im Mittel bessere Ergebnisse als diejenigen Kandidaten, bei denen Vorkenntnisse nicht vorausgesetzt werden können. Die Kandidaten mit Vorkenntnissen erzielten bessere Ergebnisse in den Leseverstehenstests als die Kandidaten ohne Vorkenntnisse zum jeweiligen Thema. Der Unterschied wird besonders pointiert hervorgehoben, wenn die Vorkenntnisse anhand von Fragen zu Schlüsselbegriffen vor dem Lesen erhoben werden. Diese Variable trennt jedoch die Gruppen auch nach dem Niveau der Deutschkenntnisse. Geringer waren die Unterschiede, wenn die Gruppen aufgrund der eigenen

Einschätzung des Textes nach dem Lesen gebildet werden. Bei dem Text "Geschwindigkeit" wiesen die Gruppen keinen Unterschied auf, was die Deutschkenntnisse betrifft. Dennoch erzielten diejenigen mit Vorkenntnissen ein besseres Ergebnis als diejenigen ohne Vorkenntnisse.

Vorkenntnisse hatten auch bei Leseverstehenstest mit eher schwach ausgeprägtem Fachbezug einen positiven Einfluss auf die Leistungen. Dass Vorkenntnisse die Leistungen im Leseverstehen beeinflussen, ist plausibel und nicht sonderlich verblüffend. Es wurde auch in mehreren Studien bestätigt. Im konkreten Fall ist diese Feststellung jedoch nicht ohne Überraschungen: Schließlich wurden Texte ausgewählt, die nur über einen sehr geringen inhaltlichen und sprachlichen Fachlichkeitsgrad verfügen. Während in anderen Studien beobachtet wurde, dass die Rolle der Vorkenntnisse mit steigendem Fachlichkeitsgrad zunimmt (siehe Kapitel 5.1), stellte sich der umgekehrte Effekt in dieser Studie nicht ein. Vorkenntnisse verloren auch bei eher geringem Fachlichkeitsgrad der Texte nicht an Bedeutung. Die Vorkenntnisse hatten vielmehr einen hoch signifikanten Einfluss auf die Leistungen im Leseverstehen.

Eine Erfassung der Vorkenntnisse ist schwierig, daher müssen mehrere Methoden eingesetzt werden. Schließlich sind Beobachtungen zur Methode zu nennen: Die Ergebnisse in den Leseverstehenstests nach Kurszugehörigkeit im Studienkolleg bzw. nach Studienziel unterscheiden sich von den Ergebnissen nach Selbstauskunft zur Vertrautheit mit dem Thema. Die meisten Kollegiaten aus den Wirtschaftskursen war mit dem Thema "Inflation" laut Selbstauskunft bereits vertraut, hinzu kamen jedoch noch einige Kollegiaten aus dem Technikkurs, für die das Thema ebenfalls nicht neu war. Ähnlich verhält es sich mit dem Thema "Geschwindigkeit": Die meisten Kollegiaten aus den Technikkursen waren mit dem Thema "Geschwindigkeitsmessung" vertraut, außerdem aber auch einige Kollegiaten aus den Wirtschaftskursen. Dies führt dazu, dass die Unterscheidung nach Studienziel/Kurszuweisung und den Vorkenntnissen nicht zu gleichen Gruppen führt. Zusätzliche Erhebungen der Vorkenntnisse sind daher sinnvoll und notwendig. Es wäre nicht angemessen gewesen, nur über die Kurszugehörigkeit und das Studienziel auf Vorkenntnisse zu einem bestimmten Thema zu schließen bzw. diese auszuschließen. Auch Wirtschaftsstudenten haben Physikkenntnisse, auch Maschinenbauer kennen volkswirtschaftliche Zusammenhänge. Bei der heterogenen Gruppe der ausländischen Studienbewerber sind die Ausbildungswege so unterschied-

lich, dass man auf eine Selbstauskunft nicht verzichten kann, wenn man Informationen über die Rolle der Fachkenntnisse beim Leseverstehen gewinnen möchte.

Fragestellung 2
Spielen die Deutschkenntnisse oder etwaige Vorkenntnisse zum Thema eine größere Rolle für die Ergebnisse in Leseverstehenstests mit geringem Fachlichkeitsgrad?

Es gibt nicht nur einen Zusammenhang zwischen Vorkenntnissen und den Ergebnissen in Leseverstehenstests mit Fachbezug, sondern auch zwischen Deutschkenntnissen und Leseverstehenstests mit Fachbezug. In diesem Kapitel wurde zunächst untersucht, ob es einen Zusammenhang zwischen dem Niveau der Deutschkenntnisse, das mit einem C-Test erhoben wurde, und Leistungen in Leseverstehenstests gab. Dies konnte anhand von Korrelationsanalysen, Streudiagrammen und Boxplots illustriert werden. Das Ergebnis entspricht den Erwartungen. Ein signifikanter Zusammenhang zwischen den Leistungen im C-Test und den Leistungen in Tests zum Leseverstehen wurde in anderen Studien ebenfalls beobachtet (Grotjahn 1995; 2002). Wenn man davon ausgeht, dass ein C-Test ein geeignetes Instrument zur Erhebung der allgemeinen (Fremd-) Sprachenkompetenz darstellt und dass die Lesefertigkeit Teil der allgemeinen Sprachkompetenz ist, war mindestens ein mittlerer Zusammenhang zwischen den Leistungen im C-Test und in den Leseverstehenstests zu erwarten. Der Zusammenhang ist dennoch erwähnenswert, deutet er doch darauf hin, dass der Fachbezug der Leseverstehenstests das Testkonstrukt nicht so beeinflusst, dass Deutschkenntnisse keine oder eine nur untergeordnete Rolle spielen.

Der Einfluss der Deutschkenntnisse auf die Leistungen in Leseverstehenstests mit Fachbezug dürfte über dem Einfluss der Vorkenntnisse liegen. Dies legen Regressionsanalysen nahe. In die Regressionsanalysen wurden folgende Variablen einbezogen: die Variablen zu den Vorkenntnissen (KURS, BEGRIFFE, BEKANNT) und die Variable zu den Deutschkenntnissen (C-TEST) als unabhängige Variablen sowie den Leseverstehenstests "Inflation" bzw. "Geschwindigkeit" als abhängigen Variablen. Die durch die Variable Deutschkenntnisse erklärte Varianz war sowohl beim Leseverstehenstest "Inflation" als auch bei "Geschwindigkeit" größer als die durch eine Variable zu den Vorkenntnissen erklärte Varianz. Im Fall von GESCHWINDIGKEIT war der "C-TEST-

Effekt" sogar stärker als der kumulierte Effekt der drei Variablen zu den Vorkenntnissen.

Die Vorhersagekraft der Deutschkompetenz und Vorkenntnisse auf Leistungen in Leseverstehenstests mit Fachbezug ist nicht besonders stark. Insgesamt war die Vorhersagekraft der Variablen C-TEST (interpretiert als "Deutschkompetenz") sowie KURS, BEGRIFFE und BEKANNT (interpretiert als "Vorkenntnisse zum Thema") aber gering. Deutschkenntnisse und Vorkenntnisse können die Ergebnisse im Leseverstehenstest "Inflation" zu 40 Prozent erklären, im Leseverstehenstest "Geschwindigkeit" zu 30 Prozent. Leseverstehenstests mit geringem Fachlichkeitsgrad enthalten also weit mehr Dimensionen als durch Ergebnisse in einem C-Test oder durch Vorkenntnisse möglicherweise erfasst werden können.

Fragestellung 3
Hängt der Einfluss der Vorkenntnisse vom Niveau der Deutschkenntnisse ab? Lassen sich sprachliche Schwellen ermitteln und beschreiben, bei denen sich der Einfluss der Vorkenntnisse zum Thema verändert?

Informationen zu diesen Fragen wurden mit einer Reihe von Varianzanalysen und Streudiagrammen mit Lowess-Regressionsgeraden erhoben. In die Varianzanalysen wurden die Ergebnisse der Kandidaten in den C-Tests, die Ergebnisse in den Leseverstehenstests mit Fachbezug sowie die Vorkenntnisse eingegeben. Dabei wurden die Kandidaten nach ihren Leistungen in den C-Tests in verschiedene Gruppen eingeteilt: Beispielsweise wurde eine Gruppe mit einem niedrigen Ergebnis im C-Test gebildet. Dann wurde diese Gruppe noch einmal nach ihren Vorkenntnissen getrennt. Gefragt wurde schließlich, ob die Ergebnisse einer Teilgruppe (z. B. niedriges Ergebnis im C-Test) in einem Leseverstehenstest signifikant höher war, wenn die Kandidaten über Vorkenntnisse verfügten.

Der Einfluss der Vorkenntnisse war nur zu einem geringen Teil abhängig vom Niveau der Deutschkenntnisse. Der Einfluss der Vorkenntnisse auf die Leistungen in Leseverstehenstests mit Fachbezug war nicht gleich bleibend. Abhängig vom Leseverstehenstest und abhängig von der Methode, mit der Vorkenntnisse erhoben wurden,

waren Wechselwirkungen zwischen den Variablen Vorkenntnisse und Deutschkenntnisse festzustellen.

Sprachliche Schwellen, bei denen sich der Einfluss der Vorkenntnisse ändert, ließen sich nicht erfassen. Obwohl sich der Einfluss der Vorkenntnisse in Abhängigkeit von dem Niveau der Deutschkenntnisse ändert, lassen sich keine ausgeprägten sprachlichen Schwellen beobachten, bei denen sich der Einfluss der Vorkenntnisse ändert. Das Niveau der Deutschkenntnisse, bei denen sich die Rolle der Vorkenntnisse veränderte, hing ab vom jeweiligen Leseverstehenstest und von der jeweiligen Methode, mit denen Vorkenntnisse erhoben wurden. Selbst allgemeine Aussagen sind nur unter Beachtung grundlegender Einschränkungen möglich. Die Gruppe der Kandidaten, bei denen Vorkenntnisse einen signifikanten Einfluss auf die Ergebnisse in Leseverstehenstests mit Fachbezug haben, ist jedoch sehr groß. Mögliche Schwellen lagen bei einem so hohen bzw. bei einem so niedrigen Ergebnis im C-Test, dass die Anzahl der Kandidaten, die zu diesen Gruppen gehörten, sehr klein war. Bei einem Vergleich unabhängiger Stichproben ist zu erwarten, dass die Varianzanalyse etwaige Mittelwertunterschiede als nicht signifikant ausweist. Verallgemeinerbare Aussagen lassen sich nicht ableiten.

Ein absichtsvoller Einsatz der sprachlichen Schwellen, bei denen sich der Einfluss der Vorkenntnisse ändert, gelang nicht. Können die vorliegenden Ergebnisse als Bestätigung der Doppelten Schwellenhypothese interpretiert werden? Die Blickrichtungen der Studien unterscheiden sich grundlegend: Claphams Studie bestand aus einer Auswertung von Ergebnissen existierender Sprachtests. Ihr Anliegen war es, an existierenden Leseverstehenstests mit geringem Fachlichkeitsgrad Informationen über den Einfluss der Vorkenntnisse zu gewinnen. In der vorliegenden Studie wurden dahingegen Leseverstehenstests unter konkreten Vorgaben erstellt. Die bewusste Herbeiführung bestimmter sprachlicher Schwellen, ab denen Vorkenntnisse keine Rolle mehr spielen, gelang dabei nicht. Lediglich die Ergebnisse einzelner Kandidaten konnten im Sinne der Doppelten Schwellenhypothese interpretiert werden. Nur in Einzelfällen war es möglich, die Ergebnisse in Leseverstehenstests durch das Zusammenspiel zwischen Deutschkenntnissen und Vorkenntnissen im Sinne der Doppelten Schwellenhypothese zu interpretieren.

Im Folgenden zeige ich mögliche Ursachen für die Ergebnisse auf und ziehe Schlussfolgerungen:

Warum war der Einfluss der Vorkenntnisse in der Studie so stark ausgeprägt?

Drei Faktoren haben meiner Ansicht nach eine Rolle gespielt: Erstens dürfte grundsätzlich gelten, dass der Einfluss der Vorkenntnisse bei Leseverstehenstests sehr wichtig ist, selbst wenn der Fachlichkeitsgrad der verwendeten Texte nicht extrem hoch ist. Leser in der Fremdsprache scheinen es leichter zu finden, Texte zu erschließen und Fragen zu Texten zu beantworten, wenn sie über Vorkenntnisse verfügen. Das Argument der Schematheorie, nach dem Textverstehen vom Vorhandensein bestehender Schemata abhängt, scheint nicht nur auf den Leseverstehensprozess in der Muttersprache, sondern auch auf die Fremdsprache zuzutreffen (siehe Alderson, 2000a: 33-48). Bedenkenswert ist zweitens der Aspekt der Motivation. Denkbar und durchaus wahrscheinlich ist, dass die Kandidaten bei Leseverstehenstests mit Bezug zu "ihrem" Studienziel motivierter waren als bei Leseverstehenstests mit einem fachfremden Thema, dass es also einen Zusammenhang gab zwischen Vorkenntnissen und Motivation und Interesse. Da die Tests für die Teilnehmer folgenlos waren, könnte der motivationale Aspekt die Auswirkungen der Vorkenntnisse erhöht haben. Erwähnt wurde in Kapitel 6.1.4 (Seite 257 ff) bereits, dass zwei der drei Variablen zu den Vorkenntnissen möglicherweise auch Deutschkenntnisse erfassen und daher den Effekt der Vorkenntnisse überzeichnen.

Wie ist die Rolle der Deutschkenntnisse für Leseverstehenstests mit Fachbezug zu bewerten?

Die Beobachtungen, die im Rahmen der vorliegenden Studie zum vergleichenden Einfluss der Vorkenntnisse und der Deutschkenntnisse gemacht wurden, stimmen mit denen anderer Studien überein. Auch in anderen Studien wurde beobachtet, dass Kenntnisse der Zielsprache für die Bearbeitung von Leseverstehenstests mit Fachbezug eine größere Rolle spielen als Vorkenntnisse zum jeweiligen Thema (siehe Kapitel 6.2.2, Seite 278 ff). Insofern entsprachen die Ergebnisse den Erwartungen.

Warum wurde die Doppelte Schwellenhypothese nicht bestätigt?

Es war nicht das vorrangige Ziel der vorliegenden Studie, die Schwellenhypothese zu bestätigen oder zu widerlegen. Ausgangspunkt der Studie war es, Informationen für die Testerstellung zu gewinnen. Es sollte in zwei konkreten Fällen beobachtet werden, wie sich der Einfluss der Vorkenntnisse in Abhängigkeit vom Niveau der Deutschkenntnisse verhält. Die Texte wurden so gewählt, dass sie sich für einen Einsatz in einem Sprachtest für den Hochschulzugang eignen. Sie hatten einen deutlich erkennbaren Fachbezug, setzten aber keine Fachkenntnisse voraus (Fachbegriffe wurden erklärt). Da der Einfluss der Vorkenntnisse groß war, ist also zu fragen, warum der Doppelte Schwelleneffekt nicht auftrat. Es gibt eine Reihe von möglichen Ursachen: Es ist möglich, dass die Schwellenhypothese sich empirisch nicht nachweisen lässt, dass sie schlicht falsch ist. Es ist weiter möglich, dass die in der Studie eingesetzten Variablen und/oder die Stichprobe eine Bestätigung der Schwellenhypothese verhinderten, dass also methodische Ursachen anzuführen sind. Auf möglichen Ursachen gehe ich ausführlicher ein:

- **Verhinderten die Texte bzw. die Leseverstehenstests eine Bestätigung der Schwellenhypothese?** Wenn das Hauptanliegen der Studie die Bestätigung der Schwellenhypothese gewesen wäre, hätten Texte mit unterschiedlichem Fachlichkeitsgrad eingesetzt werden müssen, unter anderem auch Texte mit extrem ausgeprägtem Fachlichkeitsgrad. Dies war jedoch nicht der Fall: Es wurden Texte eingesetzt, die sich für Sprachtests mit Fachbezug eignen. Es handelte sich um Texte mit eindeutigem Fachbezug, deren Fachlichkeitsgrad jedoch eher schwach ausgeprägt war. Die Auswahl der Texte wurde in Kapitel 6.1.2 (Seite 237 ff) vorgestellt und begründet. Der Einfluss der Vorkenntnisse war ausgeprägt, ich halte es daher für unwahrscheinlich, dass der Fachlichkeitsgrad der Texte so gering war, dass die Schwellenhypothese aus diesem Grunde nicht greifen konnte. Bemerkenswert ist, dass sich die Ergebnisse bei den Texten so wenig ähnelten. Während ich beim Leseverstehenstest "Inflation" geneigt bin, die Ergebnisse zumindest teilweise als in Einklang mit der Schwellenhypothese anzusehen, legen die Ergebnisse im Leseverstehenstest "Geschwindigkeit" nahe, dass die Doppelte Schwellenhypothese in

keiner Weise zutrifft. Ich sehe keinen Grund zur Annahme, dass die Auswahl der Texte den Effekt der Doppelten Schwellenhypothese beeinträchtigt haben könnte.

- **Verhinderte die Art und Weise, wie die Vorkenntnisse erhoben wurden, eine Bestätigung der Schwellenhypothese?** Vorkenntnisse wurden mit drei unterschiedlichen Methoden erhoben. Es wurde in Kapitel 6.1.4 (Seite 257 ff) dargestellt, dass die Validität von zwei Variablen möglicherweise eingeschränkt ist: Bei der Variablen KURS (Kurszugehörigkeit im Studienkolleg bzw. Studienziel) ist der genaue Bezug zum Thema der Texte unklar. Die Variable BEGRIFFE (Vertrautheit mit den Schlüsselbegriffen der Texte) erfasst möglicherweise auch Deutschkenntnisse, da Definitionen erfragt wurden. Allein bei der Variablen BEKANNT (Frage nach dem Lesen, ob das Thema bereits bekannt war) gibt es keinen Grund zur Annahme, dass Aspekte erfasst werden, die nicht repräsentativ für das Konstrukt "Vorkenntnisse zum Thema" sind, dass die Validität also eingeschränkt ist. Trotz dieser Einschränkungen gilt: Der Einfluss der Vorkenntnisse war unabhängig von der Erhebungsmethode groß, die Doppelte Schwellenhypothese wurde unabhängig von der Erhebungsmethode nicht bestätigt. Ich gehe davon aus, dass die Ergebnisse bei anderen Variablen zu Vorkenntnissen nicht grundsätzlich anders wären. Denkbar ist, dass man andere Ergebnisse gewonnen hätte, wenn man auch nach der Tiefe der Vorkenntnisse differenziert hätte. Dies wurde jedoch in der Studie nicht unternommen.

- **Verhinderte die Art und Weise, wie das Niveau der Deutschkenntnisse erhoben wurden, eine Bestätigung der Schwellenhypothese?** Die Deutschkenntnisse sind eine beachtenswerte Größe für die Schwellenhypothese. Die Wahl der C-Tests wurde in Kapitel 6.1.3 (Seite 255 ff) begründet. Unter den gegebenen Umständen (Niveau der Deutschkenntnisse sollte in kurzer Zeit erfasst werden) waren C-Tests eine angemessene Erhebungsmethode. Sie differenzieren gut zwischen den Leistungen der Teilnehmer, sie erfassen eine Reihe von sprachlichen Fertigkeiten und sie haben sich in mehreren Studien als geeignetes Instrument zur Erfassung allgemeiner Fremdsprachenkenntnisse erwiesen. Dass man ein differenzierteres Bild der Deutschkenntnisse erhalten hätte, wenn man mehrere Tests zu unterschiedlichen Fertigkeiten durchgeführt hätte, ist gleichfalls unbestritten. Ich sehe jedoch keine Anhaltspunkte dafür, dass sich die Ergebnisse bedeutsam geändert hätten.

- **Verhinderten Eigenschaften der Stichprobe, dass die Schwellenhypothese bestätigt wurde?** Hier sind zwei Punkte zu nennen: die Anzahl der Testteilnehmer und Variationsbreite der Deutschkenntnisse. Wenn die Studie darauf abgezielt hätte, die Schwellenhypothese zu bestätigen, hätte man in der Tat mehr Testteilnehmer mit besonders niedrigen und mit besonders fortgeschrittenen Deutschkenntnissen in die Stichprobe aufnehmen müssen. So argumentiert auch Clapham:

  > In my study the range of students was not wide enough for me to get a full idea of the thresholds or stages at which students at varying levels of proficiency start to use different reading processes (Clapham, 1996: 203).

  Man muss allerdings feststellen, dass diese Fragen nur von theoretischem Interesse sind; praktische Implikationen haben sie nicht. Ob ausländische Studienbewerber mit Deutschkenntnissen auf Grundstufenniveau in Leseverstehenstests von Vorkenntnissen profitieren oder nicht, ist irrelevant, da dieser Personenkreis nicht an einem Sprachtest für den Hochschulzugang teilnehmen und kein Fachstudium in deutscher Sprache aufnehmen sollte. Dass Leser in der Fremdsprache kaum von Vorkenntnissen profitieren, wenn sie nur über minimale Kenntnisse der Fremdsprache verfügen, ist darüber hinaus eine banale Feststellung. Für Sprachtests für den Hochschulzugang ist ebenfalls bedeutungslos, ob Kandidaten mit Fremdsprachenkenntnissen auf Muttersprachenniveau noch von Vorkenntnissen profitieren, da ausländische Studienbewerber typischerweise nicht über derartige Sprachkenntnisse verfügen.

- **Trifft die Doppelte Schwellenhypothese nicht zu?** Ich halte es für möglich, dass die Schwellenhypothese in der Form, wie sie von Clapham, Ridgway und anderen formuliert wurde, nicht zutrifft. Zunächst ist der Begriff Schwelle (*threshold*) unglücklich gewählt, weil er irre führende Erwartungen weckt: Bei einer Schwelle erwartet man eine plötzliche und ausgeprägte Änderung. Diese Annahme ist jedoch nicht realistisch. Selbst wenn der Grundgedanke der Doppelten Schwellenhypothese zutrifft, sollte man von einer graduellen Änderung der Rolle der Vorkenntnisse ausgehen.

  Neben die semantische Unschärfe tritt jedoch die unsichere empirische Basis, auf die die Doppelte Schwellenhypothese gestellt wurde. Die Studien von Clapham (1996) und Ridgway (1997) sind keineswegs widerspruchsfrei. Die Ergebnisse können nicht als starke Unterstützung der Doppelten Schwellenhypothese interpretiert werden. Die (Fehl-)Interpretationen dieser Studien könnten auch auf einem statistischen

Effekt beruhen: Wenn zwei (mindestens intervallskalierte) Variablen korrelieren, ist die Streuung im mittleren Bereich typischerweise größer als in den Randbereichen. Die Leistungen im Leseverstehen wurden in den Studien von Clapham, Ridgway und von mir auf einer Prozentskala gemessen, ebenso wie das Niveau der Fremdsprachenkenntnisse (mit C-Tests oder von Clapham mit einem Grammatiktest oder von Ridgway mit einem anderen Leseverstehenstest). In jedem Fall gab es eine Korrelation zwischen beiden Variablen. In einem Streudiagramm wird deutlich, dass die Streuung der Daten im mittleren Bereich groß ist (siehe die Streudiagramme in Kapitel 6.2.3, Seite 291 ff). Wenn eine dritte Variable (Vorkenntnisse) noch einen Einfluss ausübt, ist wahrscheinlich, dass dieser Effekt im mittleren Bereich am ausgeprägtesten ist, weil die Streuung dort am größten ist. Deutlich wird dieses Phänomen, wenn man besonders hohe oder niedrige Ergebnisse betrachtet. Wenn eine Kandidatin im C-Test 100 Prozent erzielte und in beiden Leseverstehenstests mit Fachbezug 100 Prozent, dann ist es unerheblich, ob sie zu einem der beiden Themen über Vorkenntnisse verfügte. Daraus ist aber nicht zwingend abzuleiten, sie benötige Vorkenntnisse nicht mehr, weil sie eine sprachliche Schwelle überschritten habe. Abzuleiten ist allein, dass die Tests in diesem Bereich nicht ausreichend differenzieren.

Die Studie zielte nicht auf eine Bestätigung der Schwellenhypothese ab, möglicherweise gab es auch Störgrößen, welche das zu erwartende Muster störten. Dennoch lassen die Ergebnisse der Studie erhebliche Zweifel an der Doppelten Schwellenhypothese aufkommen.

Lässt sich der Einfluss der Vorkenntnisse auf der Basis der Doppelten Schwellenhypothese für eine Gruppe von Testteilnehmern vorhersagen oder kontrollieren, wenn das Niveau der Sprachkompetenz sowie Schwierigkeits- bzw. Fachlichkeitsgrad der Texte in Leseverstehenstests mit Fachbezug bekannt sind?

Mit der Doppelten Schwellenhypothese wird die Erwartung verknüpft, dass man über eine gezielte Textauswahl für eine bestimmte Gruppe von Testteilnehmern den Einfluss der Vorkenntnisse vorhersagen oder kontrollieren kann (z. B. Alderson, 2000a: 104). Bei den eingesetzten Texten und Leseverstehenstests manifestierte sich das Muster der

Doppelten Schwellenhypothese nicht, obwohl es angesichts des hohen Einflusses der Vorkenntnisse zu erwarten gewesen wäre. Vor dem Hintergrund der vorgestellten Studie muss diese Frage also verneint werden. Es dürfte kaum möglich sein, das komplexe Geflecht der verschiedenen Faktoren gezielt einzusetzen. Die Faktoren, welche Ergebnisse in Leseverstehenstests beeinflussen, sind sehr unterschiedlich; unterschiedlich sind die eingesetzten Texte, die Themen und auch die Tiefe der Vorkenntnisse. Zu unterschiedlich sind möglicherweise auch die Lesefertigkeiten der Kandidaten aus unterschiedlichen Ländern.

Was bedeuten die Ergebnisse der Studie für die Diskussion um Leseverstehenstests mit Fachbezug?

Der Einfluss der Vorkenntnisse veränderte sich zwar in Abhängigkeit der Fremdsprachenkenntnisse, ein nachvollziehbares Muster war jedoch nicht abzuleiten. Es dürfte kaum erforderlich sein, die Doppelte Schwellenhypothese bei der Konstruktion und Interpretation von Sprachtests für den Hochschulzugang zu berücksichtigen. Man sollte bei Sprachtests mit Fachbezug grundsätzlich von einem ausgeprägten Einfluss der Vorkenntnisse ausgehen, selbst wenn der Fachlichkeitsgrad der Texte nicht sonderlich ausgeprägt ist. Sprachtests mit Fachbezug könnten eingesetzt werden, wenn ein Einbringen von Vorkenntnissen in Einklang mit dem Testkonstrukt und in Einklang mit der Testfunktion steht. Dies ist der Fall, wenn Aussagen über einen abgrenzbaren Bereich der Sprachverwendung getroffen werden sollen, über eine bestimmte, fachliche Sprachverwendungssituation. Dass das Testkonstrukt von Sprachtests für den Hochschulzugang ("Sprachverwendung im Fachstudium") diese Bedingung nur zum Teil erfüllt, wurde bereits diskutiert. Eine mögliche Antwort auf dieses Problem: Ein Sprachtest für den Hochschulzugang mit Fachbezug enthält nicht nur ein Thema, sondern mehrere; den Kandidaten werden mehrere Themen zur Auswahl angeboten.

Lassen sich die Ergebnisse der Studie zu Leseverstehenstests mit Fachbezug auch mit Blick auf Leseverstehenstests ohne Fachbezug, aber mit akademischem Inhalt interpretieren?

Es gibt keine verlässlichen Aussagen darüber, wie allgemein ein Thema sein muss, damit ein Einfluss der Vorkenntnisse ausgeschlossen werden kann. Die Forschungslage, über die in Kapitel 5.2 (Seite 207 ff) berichtet wurde, ist nicht eindeutig. Da die Leseverstehenstests der vorgestellten Studie über einen Fachlichkeitsgrad verfügten, der eher gering ausgeprägt war, liegt die Ansicht nahe, dass man auch bei Sprachtests ohne Fachbezug, aber mit akademischem Inhalt davon ausgehen muss, dass das Thema und die Vorkenntnisse zum Thema die Leseleistung beeinflussen. Daher sollte man in Leseverstehenstests für heterogene Gruppen von Studienbewerbern entweder mehrere Texte einsetzen (wie beim TestDaF) oder aber den Kandidaten eine Wahlmöglichkeit einräumen.

Geben die Ergebnisse der Studie zu Leseverstehenstests mit Fachbezug Hinweise zum Umgang mit einem Fachbezug in deutschen Sprachtests für den Hochschulzugang?

Wenn die mit der Doppelten Schwellenhypothese verknüpfte Erwartungen zutreffen würden, müsste man das komplizierte Beziehungsgeflecht zwischen Textschwierigkeit, Vorkenntnissen und Fremdsprachenkompetenz bei der Erstellung und Interpretation von Sprachtests mit Fachbezug berücksichtigen. Darin könnte eine Chance liegen (siehe Kapitel 5.2, Seite 225), es würde die Testerstellung aber erschweren. Die Ergebnisse der Studie legen jedoch nahe, dass man die Ergebnisse im Fall von Sprachtests für den Hochschulzugang mit Fachbezug unabhängig von der Doppelten Schwellenhypothese interpretieren sollte. Vorkenntnisse sind demnach Teil des Testkonstrukts.

Im Falle der DSH halte ich es für denkbar, dass dies mit der Testfunktion vereinbar ist. Wenn die DSH ein dezentraler Sprachtest mit einer engen Anbindung zur aufnehmenden Hochschule und zu den ausländischen Studienbewerbern bleiben soll, könnte sie durch einen Fachbezug ein eigenes, besonders Profil als hochschulnahe

Prüfung mit hochschulrelevanten Inhalten bilden. Die Erstellung von Leseverstehenstests, welche aus mehreren, kurzen Texten bestehen, dürfte sich für dezentrale Prüfungen wie der DSH jedoch kaum eignen, da die Auswahl der Texte und Items schwieriger ist. Für die DSH wäre der zweite Weg geeigneter: Das Angebot von mehreren Tests zur Auswahl. Die Anzahl hängt sicherlich von der Teilnehmergruppe ab. Wie Alderson und Urquhart feststellten, dürften zwei oder drei unterschiedliche Tests ausreichen (siehe Kapitel 5.2, Seite 207 ff). Auch wenn ein Fachbezug nur in einem Prüfungsteil oder zwei Prüfungsteilen hergestellt wird, kann dies zu einem authentischen Sprachtest führen, mit dem ein Signal für eine hochschulnahe Prüfungsvorbereitung gesetzt wird.

Für den TestDaF sehe ich diese Perspektive derzeit noch nicht. Abgesehen davon, dass der TestDaF ausdrücklich als Sprachtest ohne Fachbezug konzipiert wurde, dürfte sich die aufwändige Entwicklung von Prüfungsteilen mit Fachbezug zu unterschiedlichen Fachgebieten für einen standardisierten Sprachtest wohl erst lohnen, wenn die Zahl der Prüfungen eine andere Größenordnung erreicht.

7. Resümee

Beim TestDaF hat man den Anforderungen einer internationalen Feststellungsprüfung genüge getan und Konsequenzen gezogen: Die Nützlichkeit des TestDaF als Feststellungsprüfung wird vor allem über eine hohe Reliabilität und eine hohe Validität definiert. Für Studienbewerber aus verschiedenen Ländern mit verschiedenen Studienzielen ist es vor allem aus Gründen der Testökonomie kaum möglich, einen Fachbezug herzustellen. Auch beim Grammatiktest dürfte ein Verzicht die richtige Entscheidung sein: Die Studien gaben nur wenige Hinweise darauf, dass dieser Prüfungsteil möglicherweise unverzichtbar ist.

Auf die Weiterentwicklung des größten deutschen Sprachtests für den Hochschulzugang, der DSH, gibt es verschiedene Perspektiven. Aus Sicht des Praktikers liegt der Charme einer dezentralen Prüfung wie der DSH vor allem in der hohen Flexibilität und den Möglichkeiten, die Prüfung auf eine Zielgruppe auszurichten. Hier lohnt es sich beispielsweise über einen Fachbezug nachzudenken. Zunächst sei aber auf die Einschränkungen hingewiesen: Aus Sicht der Testtheorie stößt man bei einer dezentralen Prüfung allzu rasch an Grenzen. Die Reliabilität ist gering, damit ist auch die Validität mit Bezug auf das Testkonstrukt ("Deutschkompetenz für ein Hochschulstudium") fragwürdig (Kapitel 2.2, Seite 43 ff). Solange Prüfungskandidaten an verschiedenen Standorten unterschiedliche Ergebnisse in der DSH erzielen, ist der Nutzen der Prüfung als gering anzusehen. Hinter diesen grundsätzlichen Mängeln verblassen etwaige weitere Vorteile.

Dass Handlungsbedarf besteht, ist auch dem Fachverband Deutsch als Fremdsprache (FaDaF) hinreichend bekannt. Die eingeleiteten Maßnahmen zur Erhöhung der Reliabilität gehen in die richtige Richtung, sie greifen aus Sicht der Testtheorie aber zu kurz:

Die DSH-Rahmenordnung bleibt ein weites Dach für verschiedenartige Prüfungen, deren Verlässlichkeit nicht gesichert ist (Kapitel 2.2, Seite 43 ff) . Wenn man derartige Gedanken weiterdenkt, bleibt für die DSH nur der Weg zu einer Standardisierung – und letztlich zu einer Prüfung wie dem TestDaF. Mit dem TestDaF liegt jedoch bereits eine standardisierte Prüfung vor, welche testmethodischen Ansprüchen genügt.

Dies entspricht – wenn auch unter anderen Vorzeichen – der Situation im englischen Sprachraum: Nach der Überarbeitung wird sich der größte englische Sprachtest für den Hochschulzugang, TOEFL, kaum noch von dem zweitgrößten Test, IELTS, unterscheiden. Für den deutschen Sprachraum gibt es jedoch wenig Anlass, ein weiteres Testinstitut mit der Durchführung einer weiteren (standardisierten) Sprachprüfung für den Hochschulzugang zu beauftragen, wenn dies zu zwei weitgehend identischen Prüfungen führen würde. Im Vergleich zum englischen Sprachraum, in dem jährlich weit über eine Million Sprachprüfungen für den Hochschulzugang durchgeführt werden, liegt die Zahl ausländischer Studienanfänger in Deutschland bei ungefähr 50.000.

Derzeit sieht es so aus, als bliebe die DSH als eigenständige Prüfung erhalten. Trotz gravierender testmethodischer Mängel ist es opportun, diese Chance zu nutzen und über Möglichkeiten der Ausgestaltung und Weiterentwicklung nachzudenken. In dieser Arbeit untersuchte ich, wie man ihre Nützlichkeit erhöhen könnte. Mit Grammatik und Fachbezug wählte ich zwei Aspekte, welche die DSH vom TestDaF unterscheiden und welche die DSH gegen den derzeitigen Trend der Testmethodik positionieren würden. Unter Testentwicklern scheint sich ein Konsens gebildet zu haben, wie Sprachprüfungen für den Hochschulzugang auszusehen haben: Grammatik und Fachbezug gehören nicht dazu. Die Prüfungen mit allgemeinen Themen aus dem Hochschulumfeld bestehen aus vier Subtests zu den Fertigkeiten Lesen, Sprechen, Schreiben und Hören.

Ist der DSH-Grammatiktest dennoch ein nützlicher Prüfungsteil? Zum Testen von Grammatik liegen bislang kaum empirische Untersuchungen vor, es gibt vielmehr eine Reihe von offenen Fragen (Kapitel 3.2). Einige dieser Fragen griff ich in eigenen Studien zum DSH-Grammatiktest auf. Diese Studien ergaben jedoch kaum Hinweise darauf, dass der DSH-Grammatiktest einen wesentlichen Beitrag zur Nützlichkeit der DSH leistet: Voraussetzung für eine ausreichend hohe Reliabilität des DSH-Grammatiktests ist die Gestaltung nach einheitlichen Testmethoden-Merkmalen, welche bei der DSH bekanntlich nicht gesichert ist (Kapitel 4.1). Er ist weder als Test der all-

gemeinen Sprachkompetenz anzusehen (Kapitel 4.2), noch scheint er Informationen zu bieten, welche für die Zulassungsentscheidung von herausragender Bedeutung wären (Kapitel 4.3). Durch den DSH-Grammatiktest wird die produktive Grammatikkompetenz im Testkonstrukt abgebildet; diese Informationen werden durch andere Prüfungsteile nicht völlig erfasst. Dennoch dürfte sich das Testkonstrukt der Gesamtprüfung nur in geringem Umfang ändern. Wenn ein Sprachtest für den Hochschulzugang einen Grammatiktest enthält, scheinen sich die Kandidaten besonders auf diesen Prüfungsteil zu konzentrieren (Kapitel 4.4). Schließlich argumentierte ich in Kapitel 3.2, dass der DSH-Grammatiktest als indirekter Kompetenztest mit einem engen Bezug zum Curriculum den Charakter der DSH als Kursabschlussprüfung betont – eine mit Blick auf die Testfunktion unangemessene Vorgehensweise. Soll die DSH als Feststellungsprüfung etabliert werden, kann man auf einen Grammatiktest verzichten.

Wird ein Grammatiktest beibehalten, empfiehlt sich für die Ausgestaltung eine Kontextualisierung der Items. Auf die Verwendung von Metasprache oder auf eine Ordnung der Phänomene sollte verzichtet werden (Kapitel 4.1). Damit mögliche Auswirkungen auf die Prüfungsvorbereitung nicht zu stark sind, sollte ein Grammatiktest einen geringeren Umfang und eine geringere Gewichtung haben als andere Prüfungsteile. Dies entspricht der Ausrichtung der überarbeiteten DSH-Rahmenordnung. Eine Angliederung an das Leseverstehen bei gleichzeitigem Ergebnisausweis für beide "Unterprüfungsteile" – wie in der DSH-Rahmenordnung gefordert – ist dabei allerdings verwirrend (Kapitel 2.2, Seite 59 f).

Meiner Ansicht nach bietet Grammatik insgesamt wenig Potenzial für eine Positionierung der DSH neben dem TestDaF. Eine Beibehaltung des DSH-Grammatiktests wäre zwar ein Alleinstellungsmerkmal der DSH, der Nutzen des DSH-Grammatiktests ist jedoch nicht ausreichend hoch.

Welche Schlussfolgerungen sind mit Blick auf das zweite Thema dieser Arbeit, den Fachbezug, zu ziehen? Wegen der institutionellen Nähe und der Nähe zu den aufnehmenden Studiengängen besteht die Möglichkeit, zielgruppenspezifische Versionen der DSH anzubieten, wohingegen man auf einen speziellen Testzuschnitt beim TestDaF verzichtet. Bei der DSH kann etwa ein Fachbezug hergestellt werden. Studien zu Sprachtests mit Fachbezug legen nahe, dass Vorkenntnisse einen Einfluss auf die Ergebnisse haben, dass dieser Einfluss jedoch geringer ist als die Rolle der Fremd-

sprachenkenntnisse. Die Ergebnisse sind jedoch nicht widerspruchsfrei: häufig wurde das beste Ergebnis nicht von der Gruppe erzielt, die über Vorkenntnisse verfügt. Es wird daher spekuliert, dass der Effekt der Vorkenntnisse nicht gleichmäßig vom Niveau der Fremdsprachenkenntnisse abhängt (Kapitel 5.2). In einigen Studien wird der Eindruck vermittelt, dass es möglich sei, den Einfluss der Vorkenntnisse in Sprachtests mit Fachbezug in Abhängigkeit vom Niveau der Fremdsprachenkenntnisse zu bestimmen. Wenn dies zuträfe, müsste man diese Information für die Erstellung von Sprachtests und die Interpretation der Ergebnisse nutzen. Daher wurde die "Doppelte Schwellenhypothese" (Kapitel 5.2, Seite 220 ff) am Beispiel von Leseverstehenstests mit Fachbezug auf den Prüfstand gestellt. Der Einfluss der Fremdsprachenkenntnisse, die Rolle der Vorkenntnisse sowie die Abhängigkeit dieser Variablen voneinander wurden untersucht (Kapitel 6).

In der Studie hatten Vorkenntnisse einen signifikanten Einfluss auf die Leistungen in Leseverstehenstests mit Fachbezug. Vorkenntnisse zum Thema des Textes sind – so ist zu schlussfolgern – Teil des Testkonstrukts von Leseverstehenstests mit Fachbezug und sind daher bei der Auswahl des Themas zu berücksichtigen. Das Niveau der Deutschkenntnisse erwies sich jedoch als ein wichtigerer Faktor für die (durch den Leseverstehenstest gemessene) Leseleistung als etwaige Vorkenntnisse zum Thema.

Für die Erstellung von Sprachtests für den Hochschulzugang sind Informationen zur Abhängigkeit der Vorkenntnisse von der Sprachkompetenz aus zwei Gründen nicht produktiv nutzbar: Die mit Abstand größte Gruppe unter den ausländischen Studienbewerbern, die an Sprachprüfungen für den Hochschulzugang teilnehmen, profitiert von etwaigen Vorkenntnissen. Außerdem dürften sich kaum verlässliche Grenzen zwischen den Gruppen bestimmen lassen. Mit den eingesetzten Leseverstehenstests wurde keine Unterstützung für die "Doppelte Schwellenhypothese" gewonnen. Es war nicht möglich, sprachliche Schwellen zu beschreiben, bei denen sich der Einfluss der Vorkenntnisse zum Thema verändert. Das Abhängigkeitsverhältnis ist komplex und widersprüchlich.

Bei der Erstellung und beim Einsatz von Leseverstehenstests mit Fachlichkeitsgrad muss man davon ausgehen, dass Kandidaten mit Vorkenntnissen ein besseres Ergebnis erzielen als Kandidaten ohne Vorkenntnisse. Ob dies wünschenswert ist, hängt auch davon ab, wie zielgruppenspezifisch eine Prüfung gestaltet werden kann bzw. wie genau sich die Zielgruppe abgrenzen lässt. In diesem Sinne sehe ich den Fachbezug für die

DSH als einen Aspekt an, der wesentlich mehr Potenzial für eine zukünftige Positionierung der Prüfung bietet als ein Grammatiktest.

Wenn man eine Sprachprüfung für den Hochschulzugang mit Fachbezug konzipiert, sollten folgende Aspekte berücksichtigt werden: Prüfungsteile mit einem Fachbezug sind sinnvoll, wenn eine Streuung der Themen gewährleistet ist. Daher empfiehlt es sich, beispielsweise im Leseverstehen oder im Mündlichen Ausdruck mehrere Themen zur Auswahl anzubieten. In diesem Sinne ist eine Anbindung des Grammatiktests an das Leseverstehen kontraproduktiv. Die Empfehlungen dürften auch für Leseverstehenstests mit Bezug zu allgemeinen akademischen Inhalten gelten. Prüfungskandidaten setzen bei Sprachtests ihre Vorkenntnisse zum Thema ein.

Die Qualität der deutschen Sprachprüfungen für den Hochschulzugang ist ein wichtiger Baustein für den Studienerfolg ausländischer Studierender. Mit den Prüfungen müssen die Deutschkenntnisse der Studienbewerber transparent und zuverlässig abgebildet werden, die Prüfungen sollten dabei den Sprachgebrauch an der Hochschule reflektieren und eine sinnvolle sprachliche Studienvorbereitung fördern. Ich zeigte auf, welche Rolle die Berücksichtigung von Grammatik bzw. die Herstellung eines Fachbezugs in diesem Sinne spielen können. Vor allem für eine dezentrale Prüfung wie die DSH bietet ein Fachbezug eine Chance zur Weiterentwicklung.

Verbesserungen bei der Auswahl und der Vorbereitung ausländischer Studienbewerber und vor allem die verbesserte Förderung ausländischer Studierender im Fachstudium sind jedoch ebenfalls notwendig, damit der Studienerfolg steigt. Nur bei einer deutlich höheren Absolventenquote werden deutsche Hochschulen im Wettbewerb um internationale Studierende bestehen können. Nur wenn mehr Erfolgsgeschichten geschrieben werden, werden ausländische Studierende zu Freunden deutscher Hochschulen und zu Freunden Deutschlands.

8. Literatur

Alderson, J. C. (1981). Reaction to the Morrow paper. In Alderson, J. C.; Hughes, A. (Hrsg.). *Issues in language testing.* London: The British Council, 48-58.

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In Alderson, J. C.; Urquhart, A. H. *Reading in a foreign language.* London: Longman, 1-24.

Alderson, J. C. (1988). Testing English for specific purposes: How specific can we get? In Hughes, A. (Hrsg.). *Testing English for university study* (ELT Documents 127). London: Modern English Publications und British Council, 16-28.

Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language* 6 (2), 425-438.

Alderson, J. C. (1993). The relationship between grammar and reading in an English for Academic Purposes test battery. In Douglas, D.; Chapelle, C. (Hrsg.). *A new decade of language testing research: Selected papers from the 1990 Language testing research colloquium.* Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), 203-219.

Alderson, J. C. (1999). Reading constructs and reading assessment. In Chalhoub-Deville, M. (Hrsg.). *Issues in computer-adaptive testing of reading proficiency* (Studies in Language Testing 10). Cambridge: Cambridge University Press, 49-70.

Alderson, J. C. (2000a). *Assessing reading* (The Cambridge Language Assessment Series). Cambridge: Cambridge University Press.

Alderson, J. C. (2000b). Testing in EAP: Progress? Achievement? Proficiency? In Blue, G. M.; Milton, J.; Saville, J. (Hrsg.). *Assessing English for Academic Purposes.* Oxford u. a.: Peter Lang, 21-47.

Alderson, J. C. (2002a). Testing proficiency and achievement. In Coleman, J. A.; Grotjahn, R.; Raatz, U. (Hrsg.). *University language testing and the C-Test* (Fremdsprachen in Lehre und Forschung 31). Bochum: AKS-Verlag, 15-30.

Alderson, J. C. (2002b). *Criteria for test comparability studies.* Vortrag auf der Fachtagung "DSH und TestDaF: Politische Implikationen und wissenschaftliche Erforschung" vom 22.-23.6.2002 in Braunschweig.

Alderson, J. C.; Clapham, C.; Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research* 1 (2), 93-121.

Alderson, J. C.; Clapham, C.; Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C.; Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing* 13 (3), 281-296.

Alderson, J. C.; Urquhart, A. H. (Hrsg.). (1984). *Reading in a foreign language* (Applied Linguistics and Language Study). Harlow: Longman.

Alderson, J. C.; Urquhart, A. H. (1985a). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2 (2), 192-204.

Alderson, J. C.; Urquhart, A. H. (1985b). This test is unfair: I'm not an economist. In Hauptman, P. C.; LeBlanc, R.; Bingham-Wesche, M.: *Second language performance testing.* Ottawa: University of Ottawa Press, 25-45.

Alderson, J. C.; Wall, D. (1993). Does washback exist? *Applied Linguistics* 14 (2), 115-129.

ALTE (Association of Language Testers in Europe). (2001). *ALTE Handbuch europäischer Sprachprüfungen und Prüfungsverfahren*. Cambridge: The University of Cambridge Local Examinations Syndicate (UCLES).

Alvermann, D. E.; Hynd, C. R. (1989). Effects of prior knowledge activation modes and text structure on non-science majors' comprehension of physics. *Journal of Education Research* 83 (2), 97-102.

Arras, U.; Eckes, T.; Grotjahn, R. (2002). C-Tests im Rahmen des 'Test Deutsch als Fremdsprache' (TestDaF): Erste Forschungsergebnisse. In Grotjahn, R. (Hrsg.). *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 4). Bochum: AKS-Verlag, 175-209.

Arras, U.; Grotjahn, R. (2002). TestDaF: Aktuelle Entwicklungen. *Fremdsprache und Hochschule* 66, 65-88.

Bachman, L. F. (1990). *Fundamental considerations in language testing* (Oxford Applied Linguistics). Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing* 17 (1), 1-42.

Bachman, L. F.; Choi, I. C.; Davidson, F.; Ryan, K. (1995). *An investigation into the comparability of two tests of English as a foreign language* (Studies in Language Testing 1). Cambridge: Cambridge University Press.

Bachman, L. F.; Davidson, F.; Foulkes, J. (1993). A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL test batteries. In Chapelle, C.; Douglas, D. (Hrsg.). *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium.* Alexandria, Virginia: TESOL, 25-45.

Bachman, L. F.; Lynch, B.; Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12 (2), 238-258.

Bachman, L. F.; Palmer, A. S. (1982). The construct validation of some components of communicative competence. *TESOL Quarterly* 16 (4), 449-465.

Bachman, L. F.; Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (2003). *Multivariate Analysemethoden*. 10. Auflage. Berlin: Springer.

Bailey, K. M. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing* 13 (3), 257-279.

Baker, D. (1989). *Language testing: a critical survey and practical guide.* London: Edward Arnold.

Ballstaedt, S.-P.; Mandl, H.; Tergan, S.-O. (1982). Textverständlichkeit – Textverstehen. In Treiber, B.; Weinert, F. *Lehr-Lern-Forschung*. München, Wien und Baltimore: Urban und Schwarzenberg, 66-88.

Basiswissen Schule-Physik. (2005). Biographisches Institut und F. A. Brockhaus AG, Mannheim und Duden Paetec GmbH, Berlin.

Basiswissen Schule-Wirtschaft. (2005). Biographisches Institut und F. A. Brockhaus AG, Mannheim und Duden Paetec GmbH, Berlin.

Baur, R. S. (2001). Deutsch als Fremdsprache – Deutsch als Zweitsprache. In Helbig, G.; Götze, L.; Henrici, G.; Krumm, H.-J. (Hrsg.). *Deutsch als Fremdsprache: ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft 19.1). Berlin und New York: de Gruyter, 617-628.

Bayer, K.; Seidel, B. (1979). Verständlichkeit. *Praxis Deutsch* 36, 12-23.

Beneš, E. (1971). Fachtext, Fachstil und Fachsprache. In Institut für deutsche Sprache (Hrsg.). *Sprache und Gesellschaft* (Jahrbuch 1970 des Instituts für deutsche Sprache, Band 13). Düsseldorf: Pädagogischer Verlag Schwann, 118-132.

Bernhardt, E. B. (1991a). A psycholinguistic perspective on second language literacy. In Hulstijn, J. H.; Matter, J. F. (Hrsg.). *Reading in two languages.* AILA Review 8. Amsterdam: Free University Press, 31-44.

Bernhardt, E. B. (1991b). *Reading development in a second language: theoretical, empirical and classroom perspectives* (Second language learning series). Norwood, N. J.: Ablex Publishers.

Bernhardt, E. B. (1999). If reading is reader-based, can there be a computer adaptive test of reading? In Chalhoub-Deville, M. (Hrsg.). *Issues in computer-adaptive testing of reading proficiency* (Studies in Language Testing 10). Cambridge: Cambridge University Press, 1-10.

Bernhardt, E.; Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics* 16 (1), 15-34.

Biere, B. U. (1991). *Textverstehen und Textverständlichkeit* (Studienbibliographien Sprachwissenschaft 2). Heidelberg: Groos.

Birjandi, P.; Alavi, S. M.; Salmani-Nodoushan, M. A. (2002). *Text familiarity, reading tasks, and ESP test performance: a study on Iranian LEP and non-LEP university students* (unpublished PhD dissertation). University of Tehran. URL: www.geocities.com/nodoushan/HomePage.html (aktuell im Dezember 2005)

Black, L. (Hrsg.). (1994). *New directions in portfolio assessment: reflective practice, critical theory, and large-scale scoring.* Portsmouth: Heinemann Boyton/Cook.

Blue, G. M. (Hrsg.). (1993). *Language, learning and success, studying through English.* London: Modern English Publications and The British Council.

Bolton, S. (1997). Tests im Wandel theoretischer Prämissen. In Gardenghi, M.; O'Connell, M. *Prüfen, Testen, Bewerten im modernen Fremdsprachenunterricht* (Bayreuther Beiträge zur Glottodidaktik 6). Frankfurt am Main u. a.: Peter Lang Verlag, 17-24.

Bolton, S. (Hrsg.). (2000). *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* (Standpunkte zur Sprach- und Kulturvermittlung 8; Werkstattberichte des Goethe-Instituts). München: Goethe-Institut; Köln: Gilde Verlag.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler.* Berlin u. a.: Springer.

Bortz, J; Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.* 3. Auflage. Berlin u. a.: Springer.

Bossers, B. (1991). On thresholds, ceilings and short-circuits: The relation between L1 reading, L2 reading and L2 knowledge. In Hulstijn, J. H.; Matter, J. F. (Hrsg.). *Reading in two languages.* AILA Review 8. Amsterdam: Free University Press, 45-60.

Briest, W. (1974). Kann man Verständlichkeit messen? *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 27, 543-563.

Brindley, G. (1990). Assessing achievement in a learner-centered curriculum. In Alderson, J. C.; North, B. (Hrsg.). *Language testing in the 1990s: The communicative legacy* (Developments in English Language Teaching). London und Basingstoke: Macmillan, 153-166.

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice-Hall.

Brown, J. D. (1999). Relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing* 16 (2), 216-237.

Brown, J. D. (2001). *Using surveys in language programs* (Cambridge Language Teaching Library). Cambridge: Cambridge University Press.

Brown, J. D. (Hrsg.). (1998). *New ways of classroom assessment.* Alexandria, VA: Teachers of English to Speakers of Other Languages (TESOL).

Brown, J. D.; Hudson, T. (2002). *Criterion-referenced language testing* (Cambridge Applied Linguistics). Cambridge: Cambridge University Press.

Buhlmann, R.; Fearns, A. (2000). *Handbuch des Fachsprachenunterrichts. Unter besonderer Berücksichtigung naturwissenschaftlich-technischer Fachsprachen* (Fremdsprachenunterricht in Theorie und Praxis). 6. Auflage. Berlin: Langenscheidt.

Bußmann, H. (Hrsg.). (2002). *Lexikon der Sprachwissenschaft*. Dritte Auflage. Stuttgart: Alfred Kröner Verlag.

Butzkamm, W. (1995). Unterrichtsmethodische Problembereiche. In Bausch, K.-R.; Christ, H.; Krumm, H.-J. *Handbuch Fremdsprachenunterricht.* Tübingen und Basel: Francke Verlag, 188-194.

Cambridge University Press. (2002). *Revised CPE Handbook.* Cambridge: University of Cambridge Local Examinations Syndicate (UCLES). URL: http://www.cambridge-efl.org/support/dloads/ums.cfm (aktuell im Dezember 2005).

Canale, M. (1983). On some dimensions of language proficiency. In Oller, J. W. (Hrsg.). *Issues in language testing research.* Cambridge, Mass.: Newbury House, 254-272.

Canale, M. (1984). A communicative approach to language proficiency assessment in a minority setting. In Rivera, C. (Hrsg.). *Communicative competence approaches to language proficiency assessment: research and application.* Clevedon: Multilingual Matters, 107-122.

Canale, M.; Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1 (1), 1-47.

Carrell, P. L. (1987). Readability in ESL. *Reading in a foreign language* 4 (1), 21-40.

Carrell, P. L. (1991). Second-language reading: reading ability or language proficiency? *Applied Linguistics* 12 (2), 159-179.

Carroll, J.B. (1961). Fundamental considerations in testing for English proficiency of foreign students. In *Testing the English proficiency of foreign students*. Washington, DC: Center for Applied Linguistics, 31-40.

Casper-Hehne, W.; Koreik, U. (2002). *Perspektiven für die DSH*. Vortrag auf dem Workshop des Stifterverbands für die Deutsche Wissenschaft, der HRK und des DAAD "Wie viel Deutsch ist nötig? Sprachprüfungen für ausländische Studierende. TestDaF/DSH" vom 10. bis 11.09.2002 in Bonn.

Chalhoub-Deville, M. (Hrsg.). (1999). *Issues in computer-adaptive tests of reading* (Studies in Language Testing 10). Cambridge: Cambridge University Press.

Chalhoub-Deville, M.; Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS and TOEFL. *System* 28 (4), 523-529.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics* 19, 254-272.

Chapelle, C. A. (2001). *Computer applications in second language acquisition. Foundations for teaching, testing and research* (Cambridge Applied Linguistics). Cambridge: Cambridge University Press.

Chapelle, C. A.; Abraham; R. (1990). Cloze method: what difference does it make? *Language Testing* 7 (2), 121-145.

Chapelle, C. A.; Grabe, W.; Berns, M. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000* (TOEFL Monograph Series MS-10). Princeton, NJ: Educational Testing Service.

Chapelle, C.; Douglas, D. (Hrsg.). (1993). *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium.* Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL).

Chen, H-C.; Graves, M. F. (1995). Effects of previewing and providing background knowledge on Taiwanese college students' comprehension of American Short Stories. *TESOL Quarterly* 29 (4), 663-686.

Cheng, L.; Watanabe, Y.; Curtis, A. (Hrsg.) (2004). *Washback in language testing: research contexts and methods.* Mahwah, N. J. und London: Lawrence Erlbaum.

Chomsky, N. (1964). Degrees of grammaticalness. In Fodor, J. A.; Katz, J. J. (Hrsg.). *The structure of language: readings in the philosophy of language.* Englewood Cliffs, N. J.: Prentice-Hall, 384-389.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, Mass.: Massachusetts Institute of Technology (MIT) Press.

Chomsky, N. (1972). *Aspekte der Syntax-Theorie* (Theorie). (Übers. aus dem Amerikanischen von einem Kollektiv unter der Leitung von Ewald Lang.) Frankfurt am Main: Suhrkamp Verlag.

Clapham, C. (1996). *The development of IELTS: a study of the effect of background knowledge on reading comprehension* (Studies in Language Testing 4). Cambridge: Cambridge University Press.

Clapham, C. (2000). Assessment for academic purposes: where next? *System* 28 (4), 511-521.

Clarke, M. A. (1980). The short circuit hypothesis of ESL reading – or when language competence interferes with reading performance. *Modern Language Journal* 64, 203-209.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* 2. Auflage. Boston: Newbury House/Heinle & Heinle.

Coleman, J. A.; Grotjahn, R.; Raatz, U. (Hrsg.). (2002). *University Language Testing and the C-Test* (Fremdsprachen in Lehre und Forschung 31). Bochum: AKS-Verlag.

Crawford, James. (2004a). *Educating English learners: language diversity in the classroom.* (5. Auflage). Bilingual Education Services.

Crawford, James. (2004b) *Hard sell: why is bilingual education so unpopular with the American public?* Arizona State University: Language Policy Research Unit. URL: http://www.asu.edu/educ/epsl/LPRU/features/brief8.htm (aktuell im Dezember 2005).

Cronbach, L. J. (1971). Test validation. In Thorndike, R. L. (Hrsg.). *Educational measurement.* 2. Auflage. Washington D. C.: American Council on Education, 443-507.

Cummins, J. (1979a). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research* 49 (2), 222-251.

Cummins, J. (1979b). Cognitive-academic language proficiency, linguistic interdependence, the optimum age question, and some other matters. *Working Papers in Bilingualism* 19, 197-205.

Cummins, J. (1991). Conversational and academic language proficiency in bilingual contexts. In Hulstijn, J. H.; Matter, J. F. (Hrsg.). *Reading in two languages.* AILA Review 8. Amsterdam, 75-89.

Cummins, J. (1999). Research, ethics, and public discourse: The debate on bilingual education. *AAHE (American Associacion for Higher Education) Bulletin,* 51 (10), 3-5.

Cummins, J. (2000). BICS and CALP. Clarifying the distinction. *ERIC Digest* 438 551. URL: http://www.iteachilearn.com/cummins/bicscalp.html (aktuell im Dezember 2005).

DAAD (Deutscher Akademischer Austauschdienst). (2000). *Zweites Aktionsprogramm des DAAD zur Stärkung der internationalen Wettbewerbsfähigkeit des Studien- und Wissenschaftsstandorts Deutschland.* Bonn: DAAD.

DAAD (Deutscher Akademischer Austauschdienst). (2003). Studienverläufe im Ausländerstudium. *Pressemitteilungen des DAAD* Nr. 1 vom 8. Dezember 2003.

DAAD (Deutscher Akademischer Austauschdienst); HIS (Hochschul-Informations-System). (2003). *Ausländische Studentinnen und Wissenschaftlerinnen in Deutschland.* Sonderauswertung von DAAD und HIS zur Jahresversammlung 2003 der HRK in Dresden am 5./6. Mai 2003. Hannover: HIS.

DAAD (Deutscher Akademischer Austauschdienst); HIS (Hochschul-Informations-System). (2004). *Wissenschaft weltoffen. Daten und Fakten zur Internationalität von Studium und Forschung in Deutschland.* Bielefeld: Bertelsmann.

Davidson, F. (2000). The language tester's statistical toolbox. *System* 28 (4), 605-617.

Davies, A. (1982). Language testing. In Kinsella, V. (Hrsg.). *Surveys 1: Eight state-of-the-art articles on key areas in language teaching.* Cambridge: Cambridge University Press, 127-159.

Davies, A. (1988). Communicative language testing. In Hughes, A. (Hrsg.). *Testing English for university study* (ELT Document 127). (Ohne Ort) Modern English Publications und British Council, 5-15.

Davies, A. (1990). *Principles of language testing.* Oxford: Blackwell.

Davies, A. (1991). Language testing in the 1990s. In Alderson, J. C.; North, B. (Hrsg.). *Language Testing in the 1990s.* London und Basingstoke: Macmillan, 136-149.

Davies, A. (2001). The logic of testing Languages for Specific Purposes. *Language Testing* 18 (2), 133-147.

Davies, A.; Brown, A.; Elder, C.; Hill, K.; Lumley, T.; McNamara, T. (1999). *Dictionary of language testing* (Studies in Language Testing 7). Cambridge: Cambridge University Press.

de Jong, J. H. A. L.; Stevenson, D. K. (Hrsg.). (1990). *Individualizing the assessment of language abilities* (Multilingual Matters 59). Clevedon: Multilingual Matters.

del Pilar García Mayo, M. (2002). The effectiveness of two form-focused tasks in advanced EFL pedagogy. *International Journal of Applied Linguistics* 12 (2), 156-175.

DeMauro, G. E. (1992). Examination of the relationships among TSE, TWE and TOEFL scores. *Language Testing* 9 (2), 149-161.

DeKeyser, R. (1995). Learning second language grammar rules: an experiment with miniature linguistic system. *Studies in Second Language Acquisition* 17 (3), 379-410.

Diehl, E.; Christen, H.; Leuenberger, S.; Pelvat, I.; Studer, T. (2000). *Grammatikunterricht: Alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch*. Tübingen: Niemeyer.

Doughty, C.; Williams, J. (1998). *Focus on form in classroom second language acquisition* (Cambridge Applied Linguistics). Cambridge: Cambridge University Press.

Douglas, D. (1997). Language for specific purposes testing. In Clapham C.; Corson D. (Hrsg.). *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer, 111-119.

Douglas, D. (1998). Testing methods in context-based second language research. In Bachman, L. F.; Cohen, A. D. (Hrsg.). *Interfaces between second language acquisition and language testing research.* Cambridge: Cambridge University Press, 141-155.

Douglas, D. (2000). *Assessing language for specific purposes* (The Cambridge Language Assessment Series). Cambridge: Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: where do they come from? *Language Testing* 18 (2), 171-185.

Douglas, D.; Selinker, L. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In Chapelle, C.; Douglas, D. (Hrsg.). *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium.* Alexandria, Virginia: Teachers of English to Speakers of Other Languages (TESOL), 235-256.

Dudenredaktion (2001). *Duden – Deutsches Universalwörterbuch* (4. Auflage, CD-ROM). Mannheim: Bibliographisches Institut und F. A. Brockhaus AG.

Dudley-Evans, T.; St John, M. J. (1998). *Developments in ESP. A multi-disciplinary approach* (Cambridge Language Teaching Library). Cambridge: Cambridge University Press.

Eckerth, J. (2000). Zielsprachliche Kommunikation über Grammatik im Fremdsprachenunterricht. *Zeitschrift für Fremdsprachenforschung* 11 (1), 9-30.

Eckes, T. (2003a). Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse. *Fremdsprache und Hochschule* 69, 42-67.

Eckes, T. (2003b). *Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im TestDaF.* Vortrag auf der 31. Jahrestagung des Fachverbandes Deutsch als Fremdsprache vom 29. bis 31. Mai 2003 an der Universität Essen.

Eckes, T.; Grotjahn, R. (in Druck). Der C-Test als Ankertest für TestDaF: Analysen auf der Basis eines probabilistischen Testmodells. In Grotjahn, R. (Hrsg.) *The C-test: Theory, empirical research, applications.* Frankfurt: Lang.

Eggers, D.; Müller-Küppers, E.; Wiemer, C.; Zöllner, I. (1999). *Prüfungskurs DSH. Vorbereitung auf die Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber.* Ismaning: Hueber.

Ehlich, K. (1994). Die Lehre der deutschen Wissenschaftssprache: sprachliche Strukturen, didaktische Desiderate. In Kretzenbacher, H. L.; Weinrich, H. *Linguistik der Wissenschaftssprache* (Akademie der Wissenschaften zu Berlin, Forschungsbericht 10). Berlin und New York: de Gruyter, 325-352.

Ehlich, K. (1999). Alltägliche Wissenschaftssprache. *Info DaF* 26 (1), 3-24.

Ehlich, K. (2000). Deutsch als Wissenschaftssprache für das 21. Jahrhundert. *gfl-journal (German as a foreign language)* 1, 47-63. URL: www.gfl-journal.de/1-2000/ehlich.html (aktuell im Dezember 2005).

Ellis, N. (Hrsg.). (1994). *Implicit and explicit learning of languages.* London: Academic Press.

Ellis, R. (2001). Some thoughts on testing grammar: an SLA perspective. In Elder, C.; Brown, A.; Grove, E.; Hill, K; Iwashita, N.; Lumley, T.; McNamara, T.; O'Loughlin, K. (Hrsg.). *Experimenting with uncrtainty: Essays in honour of Alan Davies.* Cambridge: Cambridge University Press, 251-263.

Ellis, R.; Basturkmen, H.; Loewen, S. (2002). Doing focus-on-form. *System* 30 (4), 419-432.

Ercanbrack, J.; Robb, T. N. (1999). A study of the effect of direct test preparation on the TOEIC scores of Japanese university students. *Teaching English as a Second or Foreign Language – Electronic Journal (TESL-EJ)* 3 (4). URL: www-writing.berkeley.edu/TESL-EJ/ej12/a2.html (aktuell im Dezember 2005).

Erickson, M.; Molloy, J. (1983). ESP test development for engineering students. In Oller, J. (Hrsg.). *Issues in language testing research.* Rowley, MA: Newbury House, 280-288.

Erk, H. (1972). *Zur Lexik wissenschaftlicher Fachtexte: Verben – Frequenz und Verwendungsweise* (Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 4). Goethe Institut: München.

Erk, H. (1975). *Zur Lexik wissenschaftlicher Fachtexte: Substantive – Frequenz und Verwendungsweise* (Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 5). München: Hueber.

Erk, H. (1982). *Zur Lexik wissenschaftlicher Fachtexte: Adjektive, Adverbien und andere Wortarten* (Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 6). München: Hueber.

Erk, H. (1985). *Wortfamilien in wissenschaftlichen Texten* (Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 9). München: Hueber.

ETS (Educational Testing Service). (2000a). *Computer based TOEFL: score user guide.* Princeton: ETS. URL: www.toefl.org/educator/edpubs.html (aktuell im Dezember 2005).

ETS (Educational Testing Service). (2000b). *The computer based test: the new scores.* Princeton: ETS. URL: www.toefl.org/educator/edpubs.html (aktuell im Dezember 2005).

ETS (Educational Testing Service). (2003). *TOEFL test and score data summary.* Princeton: ETS. URL: www.toefl.org/educator/edpubs.html (aktuell im Dezember 2005).

ETS (Educational Testing Service). (Ohne Jahr) *TOEFL Practice Questions*. www.toefl.org/testprep/prepindx.html (aktuell im Dezember 2005).

Europarat/Rat für kulturelle Zusammenarbeit. (2001). *Gemeinsamer Europäischer Referenzrahmen für Sprachen: Lernen, Lehren, beurteilen.* Hrsg. vom Goethe-Institut Inter Nationes. Berlin, München und Wien: Langenscheidt.

FaDaF (Fachverband Deutsch als Fremdsprache). (2001). *DSH-Handbuch für Prüferinnen und Prüfer.* Münster: Fachverband Deutsch als Fremdsprache.

Farr, R. C.; Carey, R.; Tone, B. (1985). Recent theory and research into the reading process: Implications for reading assessment. In Orasanu, J. (Hrsg.). *Reading comprehension: From research to practice*. Hillsdale, N. J.: Erlbaum, 135-150.

Fischer, G. Molenaar, I. W. (Hrsg.). (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer.

Fisseni, H.-J. (1990). *Lehrbuch der psychologischen Diagnostik.* Göttingen: Hogrefe.

Flesch, R. (1949). *The art of readable writing* (25th Anniversary Edition, revised and enlarged, 1974). New York: Harper and Row.

Flesch, R. (1979). *How to write plain English. A book for lawyers and consumers.* New York: Harper and Rowe.

Flesch, R. (ohne Jahr). *How to write plain English*. URL: www.mang.canterbury.ac.nz/courseinfo/AcademicWriting/Flesch.htm (aktuell im Dezember 2005).

Fluck, H.-R. (1992). *Didaktik der Fachsprachen: Aufgaben und Arbeitsfelder, Konzepte und Perspektiven im Sprachbereich Deutsch* (Forum für Fachsprachen-Forschung 16). Tübingen: Narr.

Fluck, H.-R. (1996). *Fachsprachen. Einführung und Bibliographie* (UTB 483). 5. Auflage. Bern und München: Francke.

Fotos, S. S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute of essays on college entrance examinations? *Language Learning* 41 (3), 313-336.

Fox, J.; Graves, B.; Jennings, M.; Shohami, E. (1999). The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing* 16 (4), 426-456.

Frederiksen, J. R.; Collins, A. (1989). A systems approach to educational testing. *Educational researcher* 18 (9), 27-32.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing* 13 (1), 23-52.

Fulcher, G. (1997). Text difficulty and accessibility: Reading formulae and expert judgement. *System* 25 (4), 497-513.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics* 20 (2), 221-136.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System* 28 (4), 483-497.

Gabler Wirtschaftslexikon. (1997). 14. Auflage. Wiesbaden: Gabler.

Glienicke, S.; Katthagen, K.-M. (2003). *TestDaF – Kurs zur Prüfungsvorbereitung.* Ismaning: Hueber.

Goehte-Institut Inter Nationes. (Ohne Jahr). *Prüfungsbeschreibungen Kleines Deutsches Sprachdiplom.* URL: www.goethe.de/dll/prf/pba/kds/deindex.htm (aktuell im Dezember 2005).

Goethe-Institut Inter Nationes. (Ohne Jahr). *Prüfungsbeschreibungen Zentrale Oberstufenprüfung.* URL: www.goethe.de/dll/prf/pba/zop/deindex.htm (aktuell im Dezember 2005).

Götze, L. (2001a). Grammatiken. In Helbig, G.; Götze, L.; Henrici, G.; Krumm, H.-J. (Hrsg.). *Deutsch als Fremdsprache: ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft 19.1). Berlin und New York: de Gruyter, 1070-1078.

Götze, L. (2001b). Linguistische und didaktische Grammatik. In Helbig, G.; Götze, L.; Henrici, G.; Krumm, H.-J. (Hrsg.). *Deutsch als Fremdsprache: ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft 19.1). Berlin und New York: de Gruyter, 187-194.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly* 25 (3), 375-406.

Grabe, W. (1999). Developments in reading research and their implications for computer-adaptive reading assessment. In Chalhoub-Deville, M. (Hrsg.). *Issues in computer-adaptive tests of reading proficiency* (Studies in Language Testing 10). Cambridge: Cambridge University Press, 11-47.

Grabe, W. (2002). Reading in a second language. In Kaplan, R. B. *The Oxford handbook of applied linguistics.* Oxford: Oxford University Press, 49-59.

Grabowski, J. (1991). *Der propositionale Ansatz der Textverständlichkeit: Kohärenz, Interessantheit und Behalten.* Münster: Aschendorff.

Graefen, G. (1997). Wissenschaftssprache – ein Thema für den Deutsch-als-Fremdsprache-Unterricht? In Wolff, A.; Schleyer, W. *Fach- und Sprachunterricht: Gemeinsamkeiten und Unterschiede, Studiengänge Deutsch als Fremdsprache: Von der Theorie zur Praxis. Beiträge der 22. Jahrestagung DaF vom 2. – 28. Mai 1994 an der RWTH Aachen* (Materialien Deutsch als Fremdsprache 43). Regensburg: Fachverband Deutsch als Fremdsprache (FaDaF), 31-44.

Graefen, G. (2000). Einführung in den Gebrauch der Wissenschaftssprache. In Wolff, A.; Winters-Ohle, E. (Hrsg.). *Wie schwer ist die deutsche Sprache wirklich? Beiträge der 28. Jahrestagung DaF 2000* (Materialien Deutsch als Fremdsprache 58). Regensburg: Fachverband Deutsch als Fremdsprache (FaDaF), 191-210.

Green, P. S.; Hecht, K. (1992). Implicit and explicit grammar: An empirical study. *Applied Linguistics* 13 (2), 168-184.

Groeben, N. (1982). *Leserpsychologie: Textverständnis – Textverständlichkeit.* Münster: Aschendorff.

Gronlund, N. E. (1988). *How to construct achievement tests.* 4. Auflage. Englewood Cliffs, NF: Prentice Hall.

Grotjahn, R. (1995). Der C-Test: State of the Art. *Zeitschrift für Fremdsprachenforschung* (*ZFF*) 6 (2), 37-60.

Grotjahn, R. (2000a). Testtheorie: Grundzüge und Anwendungen in der Praxis. In Wolff A.; Tanzer H. (Hrsg.). *Sprache – Kultur – Politik* (Materialien Deutsch als Fremdsprache 53). Regensburg: Fachverband Deutsch als Fremdsprache (FaDaF), 304-341.

Grotjahn, R. (2000b). Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung des TestDaF. In:

Bolton, S. (Hrsg.). *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* (Standpunkte zur Sprach- und Kulturvermittlung 8; Werkstattberichte des Goethe-Instituts). München: Goethe-Institut; Köln: Gilde Verlag, 7-55.

Grotjahn, R. (Hrsg.). (1992). *Der C-Test: Theoretische Grundlagen und praktische Anwendungen Band 1* (Manuskripte zur Sprachlehrforschung 39/1). Bochum: Universitätsverlag Dr. N. Brockmeyer.

Grotjahn, R. (Hrsg.). (1994). *Der C-Test: Theoretische Grundlagen und praktische Anwendungen Band 2* (Manuskripte zur Sprachlehrforschung 39/2). Bochum: Universitätsverlag Dr. N. Brockmeyer.

Grotjahn, R. (Hrsg.). (1996). *Der C-Test: Theoretische Grundlagen und praktische Anwendungen Band 3* (Manuskripte zur Sprachlehrforschung 39/3). Bochum: Universitätsverlag Dr. N. Brockmeyer.

Grotjahn, R. (Hrsg.). (2002). *Der C-Test: Theoretische Grundlagen und praktische Anwendungen, Band 4* (Fremdsprachen in Lehre und Forschung 32). Bochum: AKS-Verlag.

Grotjahn, R. (2004). *Die C-Test Bibliographie.* URL: www.c-test.de (aktuell im Dezember 2005).

Grotjahn, R.; Kleppin, K. (2001). TestDaF: Stand der Entwicklung und einige Perspektiven für Forschung und Praxis. In Aguado, K.; Riemer, C. (Hrsg.). *Wege und Ziele: Zur Theorie, Empirie und Praxis der Deutschen als Fremdsprache (und anderer Fremdsprachen). Festschrift für Gert Henrici zum 60. Geburtstag* (Perspektiven Deutsch als Fremdsprache 15). Baltmannsweiler: Schneider Verlag Hohengehren, 419-434.

Gunning, R. (1952). *The technique of clear writing.* New York: McGraw-Hill.

Gutzat, B.; Kniffka, G. (2003). *Training TestDaF – Material zur Prüfungsvorbereitung: Trainingsbuch*. Berlin und München: Langenscheidt.

Hammadou, J. (1991). Interrelationships among prior knowledge inference and language proficiency in foreign language reading. *Modern Language Journal* 75, 27-38.

Hammadou, J. (2000). The impact of analogy and content knowledge on reading comprehension: What helps, what hurts. *Modern Language Journal* 84, 38-50.

Han, Y.; Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research* 2 (1), 1-23.

Harden, T.; Marsh, C. (1993). *Wieviel Grammatik braucht der Mensch?* München: iudicium Verlag.

Helbig, G. (1993). Wieviel Grammatik braucht der Mensch? In Harden, T.; Marsh, C. *Wieviel Grammatik braucht der Mensch?* München: iudicium Verlag, 19-29.

Heublein, U.; Sommer, D.; Weitz, B. (2004). *Studienverläufe im Ausländerstudium. Eine Untersuchung an vier ausgewählten Hochschulen* (Dokumentationen und Materialien). Bonn: Deutscher Akademischer Austauschdienst (DAAD).

Hinkel, E.; Fotos, S. (Hrsg.). (2001). *New perspectives on grammar teaching in second language classrooms* (ESL and Applied Linguistics Professional Series). Mahwah, New Jersey: Lawrence Erlbaum.

HIS (Hochschul-Informations-System). 2005. *Ausländische Studierende in Deutschland.*
URL: www.wissenschaft-weltoffen.de/ (aktuell im Dezember 2005).

Hock, T. S. (1990). The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In de Jong, J. H. A. L. and Stevenson, D. K. (Hrsg.). *Individualizing the assessment of language abilities.* Clevedon, PA: Multilingual Matters, 214-244.

Hoffmann, L. (1985). *Kommunikationsmittel Fachsprache. Eine Einführung.* 3. Auflage. Berlin: Akademie Verlag.

Hoffmann, L. (2001). Fachsprachen. In Helbig, G.; Götze, L.; Henrici, G.; Krumm, H.-J. (Hrsg.). *Deutsch als Fremdsprache: ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft 19.1). Berlin und New York: de Gruyter, 533-543.

Hoffmann, L.; Kalverkämper, H.; Wiegand, H. E. (Hrsg.). (1998). *Fachsprachen. Languages for Special Purposes. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. An international handbook of special-language and terminology research* (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1). Berlin und New York: de Gruyter.

Hoffmann, L.; Kalverkämper, H.; Wiegand, H. E. (Hrsg.). (1999). *Fachsprachen. Languages for Special Purposes. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. An international handbook of special-language and terminology research* (Handbücher zur Sprach- und Kommunikationswissenschaft 14.2). Berlin und New York: de Gruyter.

Howard, E. R.; Sugarman, J. (2001). Two-way immersion programs: Features and statistics. *ERIC Digest* EDO-FL-01-01, März 2001.

HRK (Hochschulrektorenkonferenz). (2000). *Rahmenordnung für die Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber (DSH).* Beschluss der HRK vom 21./22.02.2000.

HRK/KMK (Hochschulrektorenkonferenz/Kultusministerkonferenz). (2004). *Rahmenordnung über Prüfungen zum Nachweis deutscher Sprachkenntnisse.* Beschluss der HRK vom 8.6.2004.

Hudson, T. (1982). The effects of induced schemata on the "short circuit" in L2 reading: Non decoding factors in L2 reading performance. *Language Learning* 32, 1-31.

Hughes, A. (1988). Achievement and proficiency: The missing link? In Huges, A. (Hrsg.). *Testing English for university Study* (ELT Document 127). London: Modern English Publications und British Council, 36-42.

Hughes, A. (2003). *Testing for Language Teachers* (Cambridge Language Teaching Library). 2. Auflage. Cambridge: Cambridge University Press.

Hüllen, W.; Lörscher, K. (1979). Lehrbuch, Lerner und Unterrichtsdiskurs. *Unterrichtswissenschaft* 4, 313-326.

Hulstijn, J. H. (1991). How is reading in a second language related to reading in a first language? In Hulstijn, J. H.; Matter, J. F. *Reading in two languages.* AILA Review 8. Amsterdam: Free University Press.

Hung, C. M. (1990). The effects of pre-reading instruction on the comprehension of text by ESL readers. *The English Teacher* 19 (July). URL: www.melta.org.my/ET/1990 (aktuell im Dezember 2005).

Hyland, F. (2003). Focusing on form: Student engagement with teacher feedback. *System* 31 (2), 217-230.

Ingenkamp, K. (1985). *Lehrbuch der Pädagogischen Diagnostik* (Studienausgabe). 3. Auflage. Weinheim und Basel: Beltz Verlag.

International Language Testing System. (2003). *IELTS Handbook.* (Ohne Ort): British Council; IELTS Australia; University of Cambridge ESOL Examinations. URL: www.ielts.org/handbook.htm (aktuell im Dezember 2005).

Jakeman, V.; McDowell, C. (1996). *Practice tests for IELTS.* Cambridge: Cambridge University Press.

Jensen, C.; Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing* 12 (1), 99-119.

Jordan, R. R. (1997). *English for academic purposes: A guide and resource book for teachers* (Cambridge Language Teaching Library). Cambridge: Cambridge University Press.

Jung, L. (1995). *Deutsche Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber (DSH).* Ismaning: Hueber.

Kalverkämper, H. (1998). Fachsprache und Fachsprachenforschung. In Hoffmann, L.; Kalverkämper, H.; Wiegand, H. E. (Hrsg.). *Fachsprachen. Languages for Special Purposes. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. An international handbook of special-language and terminology research* (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1). Berlin und New York: de Gruyter, 48-59.

Kamil, M. L.; Samuels, S. J. (1988). Models of the reading process. In Carrell, P. L.; Devine, J.; Eskey, D. E. *Interactive approaches to second language reading* (The Cambridge Applied Linguistics Series). Cambridge: Cambridge University Press, 22-36.

Kast, B.; Neuner, G. (1994). *Zur Analyse, Begutachtung und Entwicklung von Lehrwerken für den fremdsprachlichen Deutschunterricht* (Fremdsprachenunterricht in Theorie und Praxis). Berlin u. a.: Langenscheidt.

Kenyon, D. M. (2000). Tape-mediated oral proficiency testing: Considerations in developing simulated oral proficiency interviews (SOPIs). In Bolton, S. (Hrsg.). *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* (Standpunkte zur Sprach- und Kulturvermittlung 8, Werkstattberichte des Goethe-Instituts). München: Goethe-Institut; Köln: Gilde-Verlag, 87-106.

Kieweg, W. (1999). Allgemeine Gütekriterien für Lernzielkontrollen. *Der fremdsprachliche Unterricht – Englisch* 1, 4-11.

Kniffka, G.; Üstünsöz-Beurer, D. (2001). TestDaF: Mündlicher Ausdruck. Zur Entwicklung eines kassetengesteuerten Testformats. *Fremdsprachen Lehren und Lernen* 30, 127-149.

Koh, M. Y. (1985). The role of prior knowledge in reading comprehension. *Reading in a Foreign Language* 3 (1), 375-381.

Koreik, U. (2003). DSH-TestDaF-Vergleichsstudie: Erste Ergebnisse. *FaDaF-aktuell* 2 (November 2003), 22-23.

Koreik, U. (Hrsg.). (2005). *DSH und TestDaF – eine Vergleichsstudie* (Perspektiven Deutsch als Fremdsprache, 18). Baltmannsweiler: Schneider Verlag Hohengehren.

Koreik, U.; Schimmel, D. (2002). Hörverstehenstests bei der DSH, der Feststellungsprüfung und TestDaF – eine Vergleichsstudie mit weiterführenden Überlegungen zu TestDaF und DSH. *InfoDaF* 29 (5), 409-440.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon.

Krashen, S. D. (1999). *Condemned without a trial. Bogus arguments against bilingual education.* Portsmouth, NH: Heinemann.

Krashen, S. D.; Terrell, T. D. (1983). *The natural approach: Language acquisition in the classroom.* San Francisco: Alemany Press.

Krekeler, C. (2002a). Die Grammatik fehlt! Fehlt die Grammatik? Rückwirkungsmechanismen von TestDaF und DSH. *Info DaF* 29 (5), 441-458.

Krekeler, C. (2002b). DSH und TestDaF – zwei ungleiche Sprachtests im Vergleich. *ELiSe: Essener Linguistische Skripte – elektronisch* 2 (2), 19-50. URL: www.elise.uni-essen.de/elise/elise_02_02/02krekeler_02_02.pdf (aktuell im Dezember 2005).

Krekeler, C. (2003). Der kleine Unterschied – und keine Folgen? Grammatik im TestDaF und in der DSH. *Zeitschrift für Fremdsprachenforschung* (*ZFF*) 14 (1), 107-148.

Krekeler, C. (2005). Die DSH-TestDaF-Vergleichsstudie an der Fachhochschule Konstanz. In Koreik, U. (Hrsg.). *DSH und TestDaF – eine Vergleichsstudie* (Perspektiven Deutsch als Fremdsprache, 18). Baltmannsweiler: Schneider Verlag Hohengehren, 153-185.

Kretzenbacher, H. L. (1992). *Wissenschaftssprache* (Studienbibliographien Sprachwissenschaft, 5). Heidelberg: Groos.

Kretzenbacher, H. L. (1998). Fachsprache als Wissenschaftssprache. In Hoffmann, L.; Kalverkämper, H.; Wiegand, H. E. (Hrsg.). *Fachsprachen. Languages for Special Purposes. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. An international handbook of special-language and terminology research* (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1). Berlin und New York: de Gruyter, 133-142.

Kunnan, A. J. (1999). Recent developments in language testing. *Annual Review of Applied Linguistics* 19, 235-253.

Lado, R. (1961). *Language testing. The construction and use of foreign language tests.* London: Longman.

Lado, R. (1971). *Testen im Sprachunterricht. Handbuch für die Erstellung und den Gebrauch von Leistungstests im Fremdsprachenunterricht.* Übers. von Reinhold Freudenstein. München: Max Hueber Verlag.

Langenscheidts Großwörterbuch Deutsch als Fremdsprache. (1998). Berlin und München: Langenscheidt.

Langer, I.; Schulz von Thun, F.; Tausch, R. (1999). *Sich verständlich ausdrücken.* 6. Auflage. München und Basel: Ernst Reinhardt Verlag.

Laufer, B.; Sim, D. D. (1985). Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals* 18, 405-411.

Lee, J.; Lemonnier-Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *TESOL Quarterly* 31 (4), 713-739.

Lee, W. (1998). *Prüfungen zum Nachweis deutscher Sprachkenntnisse bei ausländischen Studienbewerbern / Studienbewerberinnen (PNdS / DSH): Ihre Praxis und ihr Prüfprofil* (Materialien Deutsch als Fremdsprache 50)*.* Regensburg: Fachverband Deutsch als Fremdsprache (FaDaF).

Lexikon der Physik: in sechs Bänden. (1999). Heidelberg: Spektrum Akademischer Verlag.

Lienert, G. A.; Raatz, U. (1994). *Testaufbau und Testanalyse.* 5. Auflage. Weinheim: Beltz, Psychologie Verlags Union.

Lodewick, K. (2001). *DSH & Studienvorbereitung. Vorbereitung auf ein Studium an einer deutschsprachigen Universität.* Göttingen: Fabouda Verlag.

Lodewick, K. (2002). *TestDaF-Training. Text- und Übungsbuch zur Vorbereitung auf den TestDaF.* Göttingen: Fabouda Verlag.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In de Bot, K.; Ginsberg, R.; Kramsch, C. (Hrsg.). *Foreign language research in cross-cultural perspective.* Amsterdam: John Benjamins, 39-52.

Long, M. H.; Robinson, P. (1998). Focus on form: Theory, research, and practice. In Doughty, C.; Williams, J. *Focus on form in classroom second language acquisition* (Cambridge Applied Linguistics), 15-41.

Lutjeharms, M. (1988). *Lesen in der Fremdsprache: Versuch einer psycholinguistischen Deutung am Beispiel Deutsch als Fremdsprache* (Fremdsprachen in Lehre und Forschung 5). Bochum: AKS-Verlag.

Lutjeharms, M. (1994). Lesen in der Fremdsprache: Zum Leseprozess und zum Einsatz der Lesefertigkeit im Fremdsprachenunterricht. *Zeitschrift für Fremdsprachenforschung* 5 (2), 36-77.

McNamara, Timothy F. (1996). *Measuring second language performance*. Harlow: Longman.

McNamara, Timothy F. (1997). Performance testing. In Clapham C.; Corson C. (Hrsg.). *Encyclopedia of language and education, Bd. 7: Language testing and assessment.* Dordrecht: Kluver, 131-139.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In Wainer, H.; Braun, H. I. (Hrsg.). *Test validity.* Hillsdale, NJ: Lawrence Erlbaum, 32-45.

Messick, S. (1989). Validity. In Linn, R. L. (Hrsg.). *Educational measurement*. 3. Auflage. New York: American Council on Education/Macmillan; London: Collier Macmillan, 13-104.

Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13 (3), 241-256.

Messner, P. E.; Liu, N. (1995). The Test of English as a Foreign Language: Examination of the "cut-off scores" in US universities. *International Journal of Educational Management* 9 (2), 39-42.

Mihm, A. (1973). Sprachstatistische Kriterien zur Tauglichkeit von Lesebüchern. *Linguistik und Didaktik* 14, 117-127.

Millman, J.; Greene, J. (1989). The specification and development of tests of achievement and ability. In Linn, R. L. (Hrsg.). *Educational measurement*. 3. Auflage. New York: American Council on Education/Macmillan; London: Collier Macmillan, 335-366.

Monteiro, M. (1990). *Deutsche Fachsprache für Studenten im Ausland am Beispiel Brasiliens* (Sammlung Groos 39)*.* Heidelberg: Groos.

Morrow, K. (1979). Communicative language testing: Revolution or evolution? In Brumfit, C. J.; Johnson, K. (Hrsg.). *The communicative approach to language teaching.* Oxford: Oxford University Press, 143-159.

Morrow, K. (1986). The evaluation of tests of communicative performance. In Portal M. (Hrsg.). *Innovations in language testing.* Windsor, Berks.: National Foundation for Educational Research,
1-13.

Multhaup, U. (2002). Grammatikunterricht aus psycholinguistischer und informationsverarbeitender Sicht. In Börner, W.; Vogel, K. *Grammatik und Fremdsprachenerwerb.* Tübingen: Gunter Narr, 71-97.

Multilingual glossary of language testing terms. (1998). (Studies in Language Testing 6). Cambridge: Cambridge University Press.

Munby, J. (1978). *Communicative syllabus design.* Cambridge: Cambridge University Press.

Nebe, U. (1990). Ist Textschwierigkeit messbar? *Deutsch als Fremdsprache* 27, 350-356.

Norris, J.; Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning* 50 (3), 417-528.

Oller, J. W. (1974). Expectancy for successive elements: Key ingredients to language use. *Foreign Language Annals* 7, 443-452.

Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die Neueren Sprachen* 75, 165-174.

Oller, J. W. (1979). *Language tests at school.* London: Longman.

Oller, J. W. (1981). Language as intelligence? *Language Learning* 31, 465-492.

Oller, J. W. (1983a). "g", what is it? Response to Vollmer. In Hughes, Arthur; Porter, Don. *Current developments in language testing* (Applied Language Studies). London u. a.: Academic Press, 35-38.

Oller, J. W. (Hrsg.). (1983b). *Issues in language testing research.* Cambridge, Mass.: Newbury House.

Osman, S. (1984). Effects of prior knowledge on ESL reading. In Byong, W.-K. (Hrsg.). *Reading in Asia: The first yearbook of CCA.* Seoul: Hanyang University Press, 43-61.

Papajohn, D. (1999). The effect of topic variation in performance testing: The case of the chemistry TEACH test for international teaching assistants. *Language Testing* 16 (1), 52-81.

Peretz, A. S.; Shoham, M. (1990). Testing reading comprehension in LSP: does topic familiarity affect assessed difficulty and actual performance? *Reading in a Foreign Language* 7 (1), 447-455.

Perlmann-Balme, M. (2001). Formen und Funktionen von Leistungsmessung und –kontrolle. In Helbig, G.; Götze, L.; Henrici, G.; Krumm, H.-J. (Hrsg.). *Deutsch als Fremdsprache: ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft 19.1). Berlin und New York: de Gruyter, 994-1006.

Pichette, F.; Segalowitz, N.; Connors, K. (2003). Impact of maintaining L1 reading skills on L2 reading skill development in adults: Evidence from speakers of Serbo-Croatian learning French. *The Modern Language Journal* 87 (3), 391-403.

Pienemann, M. (1984). Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition* 6 (2), 186-214.

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypothesis. *Applied Linguistics* 10, 52-79.

Pienemann, M. (1998). *Language processing and second language development. Processability theory.* Amsterdam und Philadelphia: John Benjamins.

Pollitt, A. (1997). Rasch measurement in latent trait models. In: Clapham, C.; Corson, D. (Hrsg.). *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer, 243-253.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice* (Summer 1993), 24-31.

Projektgruppe TestDaF. (2000). TestDaF: Konzeption, Stand der Entwicklung, Perspektiven. *Zeitschrift für Fremdsprachenforschung* (*ZFF*) 11 (1), 63-82.

Purpura, J. E. (2004). *Assessing grammar* (Cambridge Language Assessment Series). Cambridge: Cambridge University Press.

Rall, M. (2001). Grammatikvermittlung. In. Helbig, G.; ;Götze, L.; Henrici, G.; Krumm, H.-J. (Hrsg.). *Deutsch als Fremdsprache: ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft 19.1). Berlin und New York: de Gruyter, 880-886.

Raupach, M. (2002). "Explizit/implizit" in psycholinguistischen Beschreibungen – eine unendliche Geschichte? In Börner, W.; Vogel, K. *Grammatik und Fremdsprachen-erwerb. Kognitive, psycholinguistische und erwerbstheoretische Perspektiven.* Tübingen: Gunter Narr Verlag, 99-117.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes* 9 (2), 109-121.

Rea-Dickins, P. (1991). What makes a grammar test communicative? In Alderson, J. C.; North, B. *Language testing in the 1990s: The communicative legacy* (Developments in English language teaching). London und Basingstoke: Macmillan, 112-131.

Rea-Dickins, P. (1997). The testing of grammar in a second language. In Clapham, C.; Corson, D. (Hrsg.). *Encyclopedia of language and education, Bd. 7: Language testing and assessment*. Dordrecht: Kluver, 87-97.

Rea-Dickins, P. (2001). Fossilisation or evolution: The case of grammar testing. In Elder, C.; Brown, A.; Grove, E.; Hill, K.; Iwashita, N.; Lumley, T.; McNamara, T.; O'Loughlin, K. (Hrsg.). *Experimenting with uncertainty. Essays in honour of Alan Davies* (Studies in Language Testing 11). Cambridge: Cambridge University Press, 22-32.

Ridgway, T. (1997). Thresholds of the background knowledge effect in foreign language reading. *Reading in a Foreign Language* 11 (1), 151-168.

Robinson, P. (1991). *ESP today: a practitioner's guide.* Hemel Hempstead: Prentice Hall International.

Rost, D. H. (1993). Assessing the different components of reading comprehension: Fact or fiction. *Language Testing* 10 (1), 79-92.

Sang, F.; Vollmer, H. J. (1978). *Allgemeine Sprachfähigkeit und Fremdsprachenerwerb: Zur Struktur von Leistungsdimensionen und linguistischer Kompetenz bei Fremdsprachenlernern.* Berlin: Max-Planck-Institut für Bildungsforschung.

Schifko, M. (2001). Prüfungen, Zertifikate, Abschlüsse als Planungskategorien für den Unterricht. In Helbig, G.; Götze, L.; Henrici, G.; Krumm, H.-J. (Hrsg.). *Deutsch als Fremdsprache: ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft 19.1). Berlin und New York: de Gruyter, 827-834.

Schröder, H. (1988). *Aspekte einer Didaktik/Methodik des fachbezogenen Fremdsprachenunterrichts* (Werkstattreihe Deutsch als Fremdsprache 20). Frankfurt am Main: Peter Lang.

Shavelson, R. J.; Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education* 18, 191-251.

Shoham, M.; Peretz, A. S.; Vorhaus, R. (1987): Reading comprehension tests: general or subject-specific? *System* 15 (1), 81-88.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1 (2), 147-170.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing* 1 (2), 202-220.

Skehan, P. (1998). *A cognitive approach to language learning* (Oxford Applied Linguistics). Oxford: Oxford University Press.

Skutnabb-Kangas, T. (1987). Erwerb einer Zweitsprache: Je früher, desto besser? – Über die Chancen sprachlicher Integration von türkischen Gastarbeiterkindern. *Deutsch lernen* 3, 3-33.

Spada, N; Lightbown, P. M. (1993). Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition* 15, 205-224.

Spolsky, B. (1973). What does it mean to know a language, or how do you get someone to perform his competence? In Oller, J. W.; Richards, J. C. (Hrsg.). *Focus on the learner.* Rowley, Mass.: Newbury House, 164-176.

Spolsky, B. (1995). *Measured words. The development of objective language testing* (Oxford Applied Linguistics). Oxford: Oxford University Press.

Stansfield, C. (1986). A history of the Test of Written English: the developmental year. *Language Testing* 3 (2), 224-234.

Stevenson, D. K. (1985). Pop validity and performance testing. In Lee, Y. P.; Fok, A. C. Y. Y.; Lord, R.; Low, G. *New directions in language testing. Papers presented at the International Symposium on Language Testing, Hong Kong* (Language Teaching Methodology Series). Oxford: Pergamon Press, 111-118.

Stiefenhöfer, H. (1985). *Lesen als Handlung – Didaktisch methodische Überlegungen zur fremdsprachlichen Lesefähigkeit.* Weinheim und Basel: Beltz.

Stiefenhöfer, H. (1995). Übungen zum Leseverstehen. In Bausch, K.-R.; Christ, H.; Krumm, H.-J. *Handbuch Fremdsprachenunterricht.* 3. Auflage. Tübingen und Basel: Francke Verlag, 246-248.

Tan, S. H. (1990). The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In de Jong, J. H. A. L.; Stevenson, D. K. (Hrsg.). *Individualizing the assessment of language abilities.* Clevedon, UK: Multilingual Matters, 214-224.

Taylor, C.; Jamieson, J.; Eignor, D.; Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL tasks* (TOEFL Research Report 61). Princeton, NJ: Educational Testing Service (ETS).

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes* 9 (2), 123-143.

Teigeler, P. (1979). Zum gegenwärtigen Stand der Verständlichkeitsforschung. *Publizistik* 24, 337-343.

TestDaF-Institut. (2001). *Informationen für Hochschulen in Deutschland.* Hagen: TestDaF-Institut/Gesamthochschule Hagen.

TestDaF-Institut. (2002). *Bewertungskriterien.* Tagungsunterlagen der FaDaF-Fachkonferenz "DSH und TestDaF: Politische Implikationen und wissenschaftliche Erforschung" am 22.-23.6.2002 in Braunschweig.

Urquhart, A. H.; Weir, C. (1998). *Reading in a second language: Process, product and practice* (Applied Linguistics and Language Study). Harlow: Longman.

Vollmer, H. J. (1983). The structure of foreign language competence. In Hughes, A.; Porter, D. (Hrsg.). *Current developments in language testing* (Applied Language Studies). London: Academic Press, 3-30.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System* 28 (4), 499-509.

Wall, D.; Alderson, J. C. (1995). Examining washback: The Sri Lankan impact study. In Cumming, A. (Hrsg.). *Validation in language testing* (Modern Languages in Practice 2). Clevedon: Multilingual Matters, 194-221.

Wall, D.; Clapham, C.; Alderson, J. C. (1994). Evaluating a placement test. *Language Testing* 11 (3), 321-344.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing* 13 (3), 319-333.

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

Weir, C. J. (1997). The testing of reading in a second language. In Clapham C.; Corson D. (Hrsg.). *Encyclopedia of language and education. Vol. 7: Language testing and assessment.* Dordrecht: Kluwer, 39-49.

Weir, C. J., Porter, D. (1994). The multi-divisible or unitary nature of reading: The language tester between Scylla and Charybdis. *Reading in a Foreign Language* 10 (2), 1-19.

Weir, C. J.; Huizhong, Y.; Yan, J. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes* (Studies in Language Testing 12). Cambridge: Cambridge University Press.

Westhoff, G. (1997). *Fertigkeit Lesen* (Fernstudieneinheit 17). München: Goethe-Institut; Berlin: Langenscheidt.

Wetzchewald, M. *Textverstehen und Textverständlichkeit – Theorie und Praxis.* URL: www.linse.uni-essen.de/esel/verständlichkeit/ (aktuell im Dezember 2005).

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Widdowson, H. G. (1979). *Explorations in applied linguistics.* Oxford: Oxford University Press.

Widdowson, H. G. (1983). *Learning purpose and language use.* Oxford: Oxford University Press.

Widdowson, H. G. (2001). Communicative language testing. In Elder, C.; Brown, A.; Grove, E.; Hill, K.; Iwashita, N.; Lumley, T.; McNamara, T.; O'Loughlin, K. (Hrsg.). *Experimenting with uncertainty. Essays in honour of Alan Davies* (Studies in Language Testing 11). Cambridge: Cambridge University Press, 12-21.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing,* 14 (1), 85-106.

Wintermann, B. (1998). Zuverlässig – Objektiv – Gültig? DSH und TestDaF – Sprachprüfungen auf dem Prüfstand. *Info DaF* 25 (1), 104-110.

Yamashita, J. (2001). Transfer of L1 reading ability to L2 reading: An elaboration of the linguistic threshold. *Studies in Language and Culture* 23 (1), 189-200.

Yamashita, J. (2004). Reading attitudes in L1 and L2, and their influence on L2 extensive reading. *Reading in a Foreign Language* 16 (1), 1-19.

Zimmermann, G. (1990). *Grammatik im Fremdsprachenunterricht der Erwachsenenbildung. Ergebnisse empirischer Untersuchungen* (Forum Sprache). Ismaning: Hueber.

Zimmermann, G.; Wißner-Kurzawa, E. (1985). *Grammatik lehren, lernen, selbstlernen. Zur Optimierung grammatikalischer Texte im Fremdsprachenunterricht* (Forum Sprache). München: Hueber.

9. Anhang

Übersicht

Grammatiktests (zu Kapitel 4)

Anhang 1: Grammatiktest "Flurbereinigung" – Prototyp aus dem DSH-Handbuch

Füllen Sie die Lücken aus, ohne die Textinformation zu verändern! Die Unterstreichungen sollen Ihnen bei der Lösung helfen:	
Fragt man nach den Ursachen der in Deutschland in den letzten Jahren verstärkt auftretenden Überschwemmungen,	Fragt man, *warum in Deutschland in den letzten Jahren verstärkt Überschwemmungen auftreten*,
so findet man einen wesentlichen Faktor in der Flurbereinigung, die in den siebziger und achtziger Jahren im Zuge der EG-Agrarpolitik vorgenommen wurde.	so findet man einen wesentlichen Faktor in der ....Flurbereinigung.
Die Flurbereinigung ist ein Eingriff in die Landschaft, der ihre natürliche Beschaffenheit nachhaltig verändert.	Die Flurbereinigung ist ein Eingriff in die Landschaft, der ....führt.
Ursprünglich lag der Flurbereinigung eine sehr sinnvolle Idee zugrunde: die durch Jahrhunderte lange Erbteilungen entstandenen kleinen und verstreuten Felder sollten zusammengelegt und flächenmäßig unter den Besitzern neu aufgeteilt werden.	Ursprünglich lag der Flurbereinigung eine sehr sinnvolle Idee zugrunde: Die kleinen und zerstreuten Felder, ... sollten zusammengelegt und flächenmäßig unter den Besitzern neu aufgeteilt werden.
Dadurch sollten größere zusammenhängende Felder entstehen, die mit Maschinen wesentlich leichter zu bearbeiten sind.	Dadurch sollten zusammenhängende Felder entstehen, die ....bearbeitet .....
Bei der Durchführung der Flurbereinigung sind Maßstäbe der rationellen Industrieproduktion auf die Landwirtschaft übertragen worden.	.... wurde, sind Maßstäbe der rationellen Industrieproduktion auf die Landwirtschaft übertragen worden.
Es blieb dabei häufig unbeachtet, dass die Lebensfähigkeit einer Landschaft von dem richtigen Funktionieren der verschiedensten Beziehungen in der Natur abhängt.	Man .... dass die Lebensfähigkeit einer Landschaft von dem richtigen Funktionieren der verschiedensten Beziehungen in der Natur abhängt.
So wurden meist der gesamte Baumbestand, Bachgehölze und Hecken entfernt, wobei auf Neuanpflanzungen verzichtet wurde.	So wurden meist der gesamte Baumbestand, Bachgehölze und Hecken entfernt, ohne dass ....
Ein günstiges Kleinklima entsteht jedoch erst durch einen guten Windschutz und durch die temperaturausgleichende und feuchtigkeits-regulierende Wirkung der Bäume und Gehölze.	Ein günstiges Kleinklima entsteht jedoch erst durch einen guten Windschutz und dadurch ....
Die natürlichen Bach- und Flussläufe wurden bei der Flurbereinigung zur Gewinnung besser zu bearbeitender Flächen häufig kanalisiert und begradigt.	Die natürlichen Bach- und Flussläufe wurden bei der Flurbereinigung häufig kanalisiert und begradigt, ...Flächen .....die man .....

Dadurch erhöht sich aber die Fließgeschwindigkeit so stark, dass in regenreichen Perioden oder zu Zeiten der Schneeschmelze mehr Wasser in die großen Flüsse strömt, als diese aufnehmen können.	Dadurch erhöht sich aber die Fließgeschwindigkeit so stark, dass in regenreichen Perioden oder zu Zeiten der Schneeschmelze mehr Wasser in die großen Flüsse strömt, als diese aufnehmen können. (Keine Änderung)

(Redaktionsgruppe im Auftrag des FaDaF, 2001, 7/4)

Anhang 2: Grammatiktest Metasprache "Teilzeitarbeit"

Vervollständigen Sie den Text auf den folgenden Seiten durch eine korrekte und sinnentsprechende Umwandlung der unterstrichenen Satzteile.

Wandel der Arbeitszeiten: Teilzeitarbeit

Teilzeitbeschäftigung ist noch überwiegend ein "weibliches" Phänomen: Die Gruppe der Teilzeitbeschäftigten besteht zu 89 % aus Frauen. Nur drei Prozent der erwerbstätigen Männer in Deutschland gehen einer Teilzeitbeschäftigung nach, und entsprechend sind nur rund 11 % aller Teilzeitbeschäftigten Männer.

1. Die mittlerweile von vielen deutschen Firmen angebotenen Teilzeitmodelle werden von Männern nur sehr schleppend angenommen. **(Relativsatz)**

Die Teilzeitmodelle, _____ _____ , werden von Männern nur sehr schleppend angenommen.

2. Dies hängt u.a. damit zusammen, dass Teilzeitarbeit auch nur zu einem Teilzeiteinkommen und zu einer Teilzeitrente führt, und dies ist für die meisten Männer noch nicht akzeptabel. **(Passiv)**

Dies hängt u.a. damit zusammen, dass Teilzeitarbeit auch nur zu einem Teilzeiteinkommen und zu einer Teilzeitrente führt, und dies

_____
_____

3. Der entscheidende Grund für die Zurückhaltung ist das Geld. Teilzeitlohn reicht weder aus, um den Lebensunterhalt eigenständig zu sichern, noch um ausreichende Renten- oder Arbeitslosengeldansprüche aufzubauen. **(2 x Nominalisierung)**

Der entscheidende Grund für die Zurückhaltung ist das Geld. Teilzeitlohn reicht weder _____, noch _____ _____ aus.

4. Wenn also die Teilzeitarbeit attraktiver gemacht werden soll, dann muss auch die Alterssicherung reformiert und von der (Vollzeit-)Beschäftigung getrennt werden. **(2 x Passiversatz)**

Wenn also die Teilzeitarbeit attraktiver gemacht werden soll, dann _____ auch die Alterssicherung _____und von der (Vollzeit) Beschäftigung _____.

5. Bemerkenswert ist die Entwicklung in den Niederlanden, wo es gelang, **(a)** die Arbeitslosenquote von zwölf auf 6,5 % zu verringern. Dies geschah unter anderem **(b)** durch eine dramatische Erhöhung der Quote der Teilzeitarbeit von 16,6 % im Jahr 1979 auf jetzt 37,4 %. **(a, Nomen b, Verbalisierung/Nebensatz)**

Bemerkenswert ist die Entwicklung in den Niederlanden, wo _____ _____ gelang. Dies geschah unter anderem

_____

6. Es ist nicht zuletzt das negative “Image”, weshalb vor allem die Männer in Deutschland den Weg in die Teilzeit nur zögerlich finden: Teilzeitarbeit trägt nach wie vor ein Stigma, sie steht im Ruf, nichts Vollwertiges zu sein. Dagegen ist es offenbar in Holland gelungen, Teilzeit gesellschaftsfähig zu machen. **(Nebensatz)**

Es ist nicht zuletzt das negative “Image”, weshalb vor allem die Männer in Deutschland den Weg in die Teilzeit nur zögerlich finden: Teilzeitarbeit trägt nach wie vor ein Stigma, sie steht im Ruf, nichts Vollwertiges zu sein,

_____, Teilzeit gesellschaftsfähig zu machen.

Textquelle: Willke, Gerhard: Die Zukunft unserer Arbeit. In: Niedersächsische Landeszentrale für politische Bildung (Hrsg.): Hannover 1998. S. 106/107

Anhang 3: Grammatiktest Ordnung "Neue Medien"

1. Teil:

Verwandeln Sie jeweils den Relativsatz in ein Partizip (Partizip I oder Partizip II) oder ein Adjektiv mit der Endung „–bar“.

Beispiel: *Die Zahl der Informationen, die in Datenbanken gespeichert sind, wächst explosionsartig.*

Lösung: *Die Zahl der in Datenbanken gespeicherten Information wächst explosionsartig.*

Einige Kritiker warnen vor den Folgen der Kommunikationsrevolution, die derzeit weltweit abläuft.

.....

Skeptiker befürchten eine Flut von Informationen, die nicht mehr kontrolliert werden kann.

.....

Der Mensch vergisst sehr schnell wieder einen großen Teil der Informationen, die er aufgenommen hat.

.....

2. Teil:

Ersetzen Sie jeweils die unterstrichenen Ausdrücke durch ein Modalverb.

Beispiel: *In der Feststellungsprüfung ist es den Studenten nicht erlaubt, elektronische Medien zu benutzen.*

Lösung: *In der Feststellungsprüfung dürfen die Studenten keine elektronischen Medien benutzen.*

Eine einzelne Kabelfaser vermag hundert Fernsehsendungen gleichzeitig zu übertragen.

.....

Immer mehr Politiker haben die Absicht, die Macht der Medienkonzerne durch Gesetze wieder zu beschränken.

.....

Es wird erwartet, dass viele Berufe ein völlig anderes Profil erhalten.

.....

3. Teil:

Verkürzen Sie jeweils die Sätze, indem Sie die Nebensätze durch Satzglieder ersetzen.

Beispiel: *Weil die Bedeutung elektronischer Medien beständig zunimmt, sollen Lehrer in Zukunft auch in Medienpädagogik ausgebildet werden.*

Lösung: *Wegen der beständigen Zunahme der Bedeutung elektronischer Medien sollen Lehrer in Zukunft auch in Medienpädagogik ausgebildet werden.*

Obwohl viele Menschen Bedenken haben, kann der Siegeszug der neuen Medien nicht mehr aufgehalten werden.

.....

Als die ersten Computer hergestellt wurden, dachte man noch nicht an die gewaltigen Folgen dieser Erfindung.

.....

Um den Einfluss ausländischer Medien zu kontrollieren, gibt es in einigen Staaten sehr strenge Gesetze.

.....

4. Teil:

Ergänzen Sie die folgenden Sätze:

Wenn alle Computer auf der Welt mit einem Schlag ausfallen würden,

.....

Je größer die Zahl der Informationen wird,

.....

Der Mensch der Zukunft muss in der Lage sein,

.....

Anhang 4: Grammatiktest "Meinungsforschung" aus der DSH-TestDaF-Vergleichsstudie

Füllen Sie die Lücken aus, ohne die Textinformation zu verändern! Die Unterstreichungen sollen Ihnen bei der Lösung helfen.

Die Meinungsforschung (Demoskopie)

Die Meinungsforschung (Demoskopie) ist ein wissenschaftliches Verfahren, mit dem die Meinung der Bevölkerung erforscht wird. Die Meinungsforschung erlebte ihren Durchbruch 1936 in den Vereinigten Staaten, als George Gallup auf der Grundlage einer repräsentativen Stichprobe den Ausgang der amerikanischen Präsidentschaftswahlen richtig vorhersagte – anders als die Zeitschrift "Literary Digest".

Beispiel: Die Redakteure der Zeitschrift "Literary Digest" waren nach Auswertung von über zwei Millionen Fragebögen davon überzeugt, dass Alfred M. Landon und nicht Franklin D. Roosevelt die Wahl gewinnen werde.	Nachdem die Redakteure der Zeitschrift "Literary Digest" über zwei Millionen Fragebögen ausgewertet hatten, waren sie davon überzeugt, dass Alfred M. Landon und nicht Franklin D. Roosevelt die Wahl gewinnen werde.
4. Gallup befragte nur wenige tausend Personen. Dadurch konnte er nicht nur das Wahlergebnis, sondern auch die zu erwartende Fehlschätzung des "Literary Digest" prognostizieren.	Gallup konnte durch . . . . nicht nur das Wahlergebnis, sondern auch die zu erwartende Fehlschätzung des "Literary Digest" prognostizieren.
5. Plötzlich wurde sich die Öffentlichkeit bewusst, dass man die Wahrscheinlichkeitsrechnung auf die politische Meinungsbildung anwenden konnte.	Plötzlich wurde sich die Öffentlichkeit bewusst, dass sich. . . . .
Mit diesem spektakulären Erfolg setzte der Siegeszug der Demoskopie ein.	Mit diesem spektakulären Erfolg setzte der Siegeszug der Demoskopie ein.
6. Der amerikanische Präsident Franklin D. Roosevelt beispielsweise ließ sich ab Beginn der vierziger Jahre von Hadley Cantril, einem an der Universität Princeton lehrenden Meinungsforscher, beraten.	Der amerikanische Präsident Franklin D. Roosevelt beispielsweise ließ sich ab Beginn der vierziger Jahre von Hadley Cantril, einem Meinungsforscher, ... ,beraten.
7. Trotz einiger Rückschläge - nicht zuletzt Gallups Fehlprognose bei den Präsidentschaftswahlen von 1948 – wurde die politische Umfrageforschung ein Teil des politischen Lebens.	Obwohl. . . . . - nicht zuletzt Gallups Fehlprognose bei den Präsidentschaftswahlen von 1948 – wurde die politische Umfrageforschung ein Teil des politischen Lebens.
Die Kritik an der Demoskopie ist dabei so alt wie die Demoskopie selbst.	Die Kritik an der Demoskopie ist dabei so alt wie die Demoskopie selbst.
8. Doch trotz fortdauernder Kritik kann man die Meinungsforschung aus dem politischen Leben nicht mehr wegdenken.	Doch trotz fortdauernder Kritik . . . . nicht mehr wegzudenken.
9. Vor allem in Wahljahren erlangt die Meinungsforschung eine besonders große Bedeutung und die Daten werden in immer kürzeren Abständen veröffentlicht.	Vor allem in Wahljahren erlangt die Meinungsforschung eine besonders große Bedeutung und . . . . in immer kürzeren Abständen.
10. Die Herausforderungen für die Meinungsforscher sind ebenfalls gewachsen - angesichts einer Wählerschaft, die sich erst kurz vor dem Wahltermin entscheidet.	Die Herausforderungen für die Meinungsforscher sind ebenfalls gewachsen - angesichts einer Wählerschaft, die erst kurz vor dem Wahltermin eine . . . . . .

11. So antworteten in einer Wahltagsbefragung 1998 16 Prozent, sich erst am Wahlsonntag für eine Partei entschieden zu haben.	So antworteten in einer Wahltagsbefragung 1998 16 Prozent, dass . . . .
12. Umso größer ist das Verlangen der Politiker nach Daten über die Bevölkerung, die in regelmäßigen Intervallen erhoben werden.	Umso größer ist das Verlangen der Politiker nach . . . . Daten über die Bevölkerung.
13. Die Meinungsforschung hat sich – trotz des ihr weiterhin entgegengebrachten Misstrauens - fest im politischen System Deutschlands etabliert und sie gewinnt weiter an Bedeutung.	Die Meinungsforschung hat sich - trotz des Misstrauens, ... - fest im politischen System Deutschlands etabliert und sie gewinnt weiter an Bedeutung.

Nach: Gallus, Alexander (2002). "Demoskopie in Zeiten des Wahlkampfs". Aus Politik und Zeitgeschichte, Bd. 15-16, 2

Leseverstehenstests mit Fachbezug (zu Kapitel 6)

Anhang 5: Leseverstehenstests: Kurzfragebogen zur Erfassung der Vorkenntnisse

Fragen
Name: ..... 1. Welchen Kurs besuchen Sie? (z. B. Wirtschaftskurs) ..... 2. Haben Sie bereits an einer Universität studiert? ❒ ja ❒ nein Wenn ja: was? ..... 3. Haben Sie bereits einen Beruf ausgeübt? ❒ ja ❒ nein Wenn ja: welchen Beruf? ..... 4. Was möchten Sie studieren? ..... 5. Was ist ein "**Laser**"? ..... 6. Was versteht man unter "**Radar**"? ..... 7. Was spricht man von "**Inflation**"? ..... 8. Wann liegt eine "**Deflation**" vor? .....

Anhang 6: C-Tests zur Erhebung der Sprachkenntnisse

Lückentexte

Füllen Sie die Lücken aus!

Bedeutung der Lesefähigkeit

Die Lesefähigkeit trägt ihren Wert natürlich in sich, hat aber auch ökonomische Auswirkungen. Erwac_____ Leser, d_____ besser le_____ als d_____ Durchschnitt, üb_____ mit größ_____ Wahrscheinlichkeit gutbe_____ Berufe a_____. Die wach_____ Spezialisierung i_____ der Gesell_____ erfordert me_____ Bildung, ei_____ Forderung, d_____ vor al_____ an d_____ Schulen geri_____ wird. Du_____ die erhö_____ Anforderungen a_____ das Bildungsniveau, die heute in den westlichen Gesellschaften gestellt werden, ist die Lesefähigkeit des Einzelnen immer wichtiger geworden.

Naturkatastrophen

Neben den plötzlich auftretenden Naturkatastrophen gibt es natürliche Risiken, die kontinuierlich vorhanden und schwer erkennbar sind: etwa die natürlich vorkommende Radioaktivität oder natürliche toxische Metallvorkommen in der Umwelt. Zu_____ können ein_____ natürliche Ris_____ durch d_____ Eingriffe d_____ Menschen verschlimmert wer_____: etwa Überschw_____ aufgrund d_____ Zerstörung v_____ Wäldern. F_____ die Erfor_____ dieser Gefa_____ sind des_____ die grundl_____ Erkenntnisse d_____ Umweltwissenschaften v_____ zentraler Bede_____. Die schwe_____ Risiken du_____ Naturkatastrophen best_____ in den wirtschaftlich noch wenig entwickelten Staaten. Dies liegt teils an den klimatischen Bedingungen der Tropen, teils an der Lage innerhalb geologischer Schwäche- oder Gefahrenzonen und schließlich an der noch gering ausgebauten Infrastruktur bezüglich Schutzmaßnahmen für Mensch und Umwelt.

Anhang 7: Text "Geschwindigkeit"

Lesen Sie den Text und bearbeiten Sie die folgenden Aufgaben.

Geschwindigkeitsmessung

Die Geschwindigkeit eines Körpers kann allgemein berechnet werden mit den Gleichungen

Dabei bedeuten: v Geschwindigkeit, s zurückgelegter Weg, t benötige Zeit.

$$v = \frac{s}{t} \quad \text{oder} \quad \Delta v = \frac{\Delta s}{\Delta t}$$

Bei der Geschwindigkeit ist zwischen der Durchschnittsgeschwindigkeit und der Augenblicksgeschwindigkeit (Momentangeschwindigkeit) zu unterscheiden. Die Durchschnittsgeschwindigkeit gibt an, wie groß die mittlere Geschwindigkeit längs einer Strecke ist, die ein Körper in einer bestimmten Zeit zurücklegt. Die Augenblicksgeschwindigkeit gibt die Geschwindigkeit zu einem bestimmten Zeitpunkt an. Je nach der Art der Messung erhält man entweder Durchschnittsgeschwindigkeiten oder näherungsweise Augenblicksgeschwindigkeiten. Die Zeitmessung kann mit einer Stoppuhr, aber auch elektrisch erfolgen.

Die Geschwindigkeit von Fahrzeugen kann man z. B. mithilfe von Induktionsschleifen messen. Dazu werden in der Fahrbahn zwei Induktionsschleifen (Leiterschleifen) verlegt. Die Fahrzeit für den Weg von Schleife zu Schleife wird gemessen. Genutzt wird dabei folgender Effekt: Wird eine solche Induktionsschleife von einem Strom durchflossen, so bildet sich um sie ein magnetisches Feld. Fährt ein Auto über eine solche Schleife, so wird das Magnetfeld beeinflusst. Die Magnetfeldänderung bewirkt nach dem Induktionsgesetz eine Induktionsspannung und einen Induktionsstrom, der registriert werden kann. Mit zwei Induktionsschleifen kann der zeitliche Abstand zwischen den beiden Magnetfeldänderungen registriert werden. Aus der gemessenen Zeit und dem Abstand der Schleifen ergibt sich die Geschwindigkeit. Bei größerem Abstand der Schleifen erhält man eine Durchschnittsgeschwindigkeit, bei kleinem Abstand näherungsweise die Augenblicksgeschwindigkeit des Autos. Eine solche Anordnung von zwei Induktionsschleifen kann auch mit einer automatischen Kamera gekoppelt werden, die bei Geschwindigkeitsüberschreitungen ausgelöst wird und das Fahrzeug des Verkehrssünders aufnimmt.

Eine weit verbreitete Methode, die z. B. auch von der Polizei bei Geschwindigkeitskontrollen zunehmend angewendet wird, ist die Lasermessung. Dabei werden von einem Gerät (Laserpistole) Lichtimpulse ausgesandt. Diese werden vom heranfahrenden Fahrzeug reflektiert

und vom Empfänger, der in die Laserpistole integriert ist, aufgenommen. Die Aussendung der Lichtimpulse erfolgt in sehr kurzen zeitlichen Abständen von ca. 0,02 Sekunden. Aus der Zeit, die ein Lichtimpuls für den Hin- und Rückweg benötigt, kann die Entfernung ermittelt werden, da die Geschwindigkeit der Lichtimpulse bekannt ist. Sie ist gleich der Lichtgeschwindigkeit, beträgt also etwa 300.000 km/s. Aus der Entfernungsänderung von einem Lichtimpuls zum nächsten erhält man den zurückgelegten Weg. Die Zeit, die zwischen der Aussendung von zwei Impulsen vergeht, ist die zugehörige Zeit. Damit kann die Geschwindigkeit des angepeilten Fahrzeuges berechnet und direkt angezeigt werden. Ähnlich wie bei Induktionsschleifen kann das Gerät auch mit einer automatischen Kamera gekoppelt werden, die bei Geschwindigkeitsüberschreitungen ausgelöst wird und das Fahrzeug des Verkehrssünders aufnimmt.

Ein anderes Verfahren ist die Radarmessung, bei der ein anderer physikalischer Effekt, der DOPPLER-Effekt, genutzt wird. Bei einem Radarmessgerät werden elektromagnetische Wellen kurzer Wellenlänge abgestrahlt, vom Fahrzeug reflektiert und wieder empfangen. Bewegt sich das Fahrzeug vom Messgerät weg oder auf das Messgerät zu, so tritt eine Frequenzänderung auf, die ein Maß für die Geschwindigkeit des Fahrzeuges ist.

Quelle: Basiswissen Schule-Physik. (2005). Biographisches Institut und F. A. Brockhaus AG, Mannheim und Duden Paetec GmbH, Berlin. *453 Wörter*.

Anhang 8: Leseverstehenstest "Geschwindigkeit" - Items

Beantworten Sie die folgenden Fragen in Stichworten.

1) Welche Messgrößen werden benötigt, um die Geschwindigkeit eines Fahrzeugs zu bestimmen?

. . . . . .

2) Wie entsteht Induktionsstrom?

. . . . . .

3) Was passiert, wenn Strom durch eine Induktionsschleife fließt?

. . . . . .

4) Was passiert, wenn ein Auto über ein magnetisches Feld fährt?

. . . . . .

5) Warum wird die Geschwindigkeitsmessung häufig mit einer Kamera gekoppelt?

. . . . . .

6) Wie kann man näherungsweise eine Momentangeschwindigkeit messen?

. . . . . .

7) Was ist die Aufgabe des „Empfängers“ einer Laserpistole? (Zeile 51)

. . . . . .

8) Wie wird die Strecke bei der Geschwindigkeitsmessung mittels Laser bestimmt?

. . . . . .

9) Wann ist bei einer Radarmessung eine Veränderung der Frequenz festzustellen?

. . . . . .

10) Wie viele Messungen sind zur Bestimmung der Geschwindigkeit mittels Laser notwendig?

. . . . . .

11) Wessen Frequenz ändert sich? (Zeile 77)

. . . . . .

12) Worauf bezieht sich „dazu“? (Zeile 21)

. . . . . .

13) Worauf bezieht sich „sie“? (Zeile 26)

. . . . . .

Füllen Sie die Lücken aus.

a) Die mittlere Geschwindigkeit während der gesamten Fahrt bezeichnet man auch als .....

b) Die Erzeugung einer elektrischen Spannung mit Hilfe veränderlicher magnetischer Felder nennt man ..... .

c) Verändert sich ..... der elektromagnetischen Wellen, kann die Geschwindigkeit eines Körpers bestimmt werden.

Anhang 9: Vergleichstext "Radar"

Radar, *Radio Detection and Ranging*, Verfahren zur Entdeckung und Positionsbestimmung von festen und bewegten Objekten mit Hilfe elektromagnetischer Wellen.

Das Radar arbeitet nach dem Prinzip eines Echolots: Der Radarsender strahlt elektromagnetische Wellen im mm- bis m-Bereich aus, deren Reflexionen ausgewertet werden. Der Ort eines vom Radar erfassten Objekts wird aus der Laufzeit und der Richtung des Echos bestimmt; unter Ausnutzung des Doppler-Effektes kann die Relativgeschwindigkeit zwischen Radargerät und Zielobjekt berechnet werden. Gegenüber optischen oder akustischen Ortungsverfahren besteht der Vorteil der Radartechnik im hohen Durchdringungsvermögen der Funkwellen und ihrer größeren Reichweite.

Beim Impulsradar werden die Funkwellen in Form kurzer Impulse (0,05-1µs) abgestrahlt. Die Vorteile dieses Betriebsregimes sind neben einer Energieersparnis die einfache Bestimmung der Laufzeit der Impulse und die Möglichkeit der Doppelnutzung der Radarantenne zum Senden und Empfangen. Die im Muttergenerator erzeugten Impulse werden gleichzeitig über den Modulator an den Sender und als Steuerimpuls an das Sichtgerät (bzw. die Auswerteelektronik) gegeben. Der Duplexer (Sende-Empfangs-Weiche) verhindert ein Übersprechen der Suchimpulse auf den Empfänger. Wird als Sichtgerät eine Elektronenstrahlröhre verwendet, so steuert der Muttergenerator die Zeitablenkung und der im Empfänger aufbereitete Echoimpuls die Vertikalablenkung oder die Intensität des Elektronenstrahls. Auf dem Bildschirm erscheint ein Zacken oder Leuchtfleck, dessen Lage durch die Laufzeit des Impulses bestimmt ist und somit der Entfernung zum reflektierenden Objektes entspricht.

Beim Dauerstrichradar (CW-Radar, Continuous-Wave-Radar) wird kontinuierlich gesendet, ein durch Modulation der Trägerfrequenz aufgeprägtes Signal gestattet die Laufzeitbestimmung.

Beim SLR-Radar (*side looking radar*) mit realer Apertur wird die Bildszene seitwärts blickend quer zur Flugrichtung streifenweise durch das von einer Antenne ausgestrahlte Radarsignal beleuchtet. Die unterschiedlichen Laufzeiten der vom Bodenprofil zurückgeworfenen Radarechos werden dabei auf elektronischem Wege zu einem Echobild der überflogenen Landschaft aufgebaut. Beim SAR-Radar (*synthetic aperture radar*) werden Richtstrahlen mit weitem Öffnungswinkel verwendet, wodurch der in der Flugbahn bewegte Sensor vom gleichen Objekt eine Vielzahl von Echos empfängt, und zwar solange, wie das Objekt vom Richtstrahl überstrichen wird. Amplitude und Phase der vom Objekt reflektierten Signale werden in dieser Zeit elektronisch gespeichert. Die während dieser Zeit zurückgelegte Flugstrecke entspricht der

synthetischen Apertur des Aufnahmesystems. Im Gegensatz zum eben beschriebenen Primär-Radar sendet beim Sekundär-Radar ein im Zielobjekt befindlicher Sender als Reaktion auf die Suchsignale eine eigene Kennung. Somit ist eine Identifizierung des erfassten Objektes möglich.

Neben der militärischen Nutzung kommt die Radar-Technik für viele zivile Aufgaben wie Flugsicherung und Navigation, Überwachung des Luft- und erdnahen Weltraums oder als Wetter- und Verkehrsradar (Radarfalle) zum Einsatz, häufig als integraler Bestandteil eines umfassenderen Informationssystems.

Quelle: Lexikon der Physik: in sechs Bänden. *1999. Band 4. Heidelberg: Spektrum Akademischer Verlag.*

Anhang 10: Text "Inflation"

Lesen Sie den Text und bearbeiten Sie die folgenden Aufgaben.

Inflation und Deflation

Inflation ist die über einen längeren Zeitraum festzustellende Erhöhung des Preisniveaus. Dabei ist die Zunahme der Preise im Durchschnitt, bei Beachtung der Gewichtung der Waren, von Bedeutung und nicht die Erhöhung einzelner Preise. Man misst die Inflation mit dem Preisindex für die Lebenshaltung aller privaten Haushalte. Die Ermittlung des Preisindexes erfolgt durch Feststellung der Entwicklung der Ausgaben des Durchschnittshaushaltes für einen repräsentativen Warenkorb einer Periode. Beispiel: Ausgaben für einen Warenkorb im Jahre x: 12000,-- DM; Ausgaben für einen Warenkorb im Jahre y: 12300,-- DM. Der Preisindex beträgt demnach:

$$\frac{12300}{12000} \times 100 = 102{,}5 = 2{,}5\,\%\,\text{Inflationsrate}$$

Die Preissteigerungs- bzw. Inflationsrate gibt an, um wie viel Prozent der repräsentative Warenkorb im Vergleich zum Vorjahr teurer geworden ist. Die Inflationsrate wird monatlich veröffentlicht. Das verwendete Basisjahr wird für einen längeren Zeitraum (meist 5 Jahre) beibehalten. Gegenwärtig wird als Basisjahr 1995 zugrundegelegt. Der Preisindex für die Lebenshaltung aller privaten Haushalte betrug 1998 = 104,3 %, d. h., gegenüber dem Jahre 1995 (= 100 %) beträgt die Preissteigerung 4,3 % bzw. die Inflationsrate macht 4,3 % aus. Der Begriff „Warenkorb“ ist die Bezeichnung sämtlicher für die Berechnung des Preisindexes ausgewählter Güter, d. h. Waren (Nahrungsmittel, Kleidung, Tabakwaren, Hausrat usw.), Leistungen (Verkehr, Versicherung) und Mietwohnungen, die als repräsentativ gelten. In der Regel werden nur Preissteigerungen von einer gewissen Dauer als Inflation bezeichnet. Damit werden saisonale Preissteigerungen nicht berücksichtigt.

Der Warenkorb für den Preisindex zeigt die Gewichtung des Warenkorbes für alle privaten Haushalte. Mit 27,5 % nehmen Wohnung und Energie den Hauptanteil ein, gefolgt von den Verkehrsaufwendungen, z.B. Auto und den Nahrungsmitteln. Durch solche Veränderungen in den Verbrauchsgewohnheiten (Konsumverhalten) und verbesserte Güterqualitäten bzw. neue Güter mit längerer Lebensdauer ist eine Neubestimmung des Warenkorbes immer nach einigen Jahren notwendig.

Preiserhöhungen führen zu einer permanenten Geldentwertung, d. h., wie im Beispiel gezeigt, wenn man im Jahre y für eine Summe von Waren mehr Geld ausgeben muss als im Jahre x, dann ist der Geldwert gesunken. Letztlich kommt das in einer sinkenden Kaufkraft der

Währungen zum Ausdruck und damit in einer Senkung der Reallöhne, die durch Lohnkämpfe wieder ausgeglichen werden können. Im internationalen Vergleich liegt die deutsche Inflationsrate in den 90iger Jahren unter dem Durchschnitt der meisten europäischen Länder. Auch gegenüber den USA ist der Preisanstieg geringer ausgefallen.

Gegenüber der Inflation stellt die Deflation einen absoluten Rückgang des Preisniveaus dar. Bei der Deflation handelt es sich um eine Unterversorgung der Volkswirtschaft mit Zahlungsmitteln, die zu einem Sinken der Preise führt bzw. zu einer Erhöhung des Tauschwertes des Geldes. Die Ursache der Deflation kann in zu geringer Verschuldungsbereitschaft der Produzenten oder im Horten von Bargeld liegen. Eine Deflation kann durch währungs- und kreditpolitische Maßnahmen, wie Einschränkung der Kreditvergabe und damit der umlaufenden Geldmenge sowie verschärfter Anforderungen an die Kreditwürdigkeit bewusst herbeigeführt werden, um einer drohenden Inflation entgegenzuwirken. Eine Deflation wirkt produktions-, beschäftigungs- und einkommensmindernd. In Industrieländern hat die Deflation nur noch historische Bedeutung. Der Begriff Deflation ist nicht zu verwechseln mit der Disinflation, die die Abnahme der Inflationsrate bezeichnet.

Quelle: Basiswissen Schule-Wirtschaft. (2005). Biographisches Institut und F. A. Brockhaus AG, Mannheim und Duden Paetec GmbH, Berlin. *486 Wörter*.

Anhang 11: Leseverstehenstest "Inflation" – Items

Beantworten Sie die folgenden Fragen mit Stichworten.

1) Was versteht man unter einer Inflation?

   .....

2) Was wird mit der Zahl 4,3 % genau ausgedrückt? (Zeile 27)

   .....

3) Welche Auswirkungen hat eine Inflation normalerweise auf die Kaufkraft?

   .....

4) Welche Auswirkungen hat eine Deflation normalerweise auf die Lohnentwicklung?

   .....

5) Wie verhält sich die umlaufende Geldmenge bei einer Deflation?

   .....

6) Wann könnte der Staat ein Interesse an einer Deflation haben?

   .....

7) Welche Rolle spielt die Deflation derzeit in Europa?

   .....

8) Worauf bezieht sich „sämtlicher“? (Zeile 29)

   .....

9) Worauf bezieht sich „das“? (Zeile 52)

   .....

10) Worauf bezieht sich das erste und worauf das zweite „die“? (Zeile 79)
    1. „die“..... 2. „die“ .....

Füllen Sie die Lücken aus.

a) Die Inflation wird mithilfe ..... ermittelt.
b) Wenn der Preisindex für die Lebenshaltung aller privaten Haushalte gesunken ist, spricht man von einer .....
c) Der Warenkorb wird einmal mit den Preisen des ....., zum anderen mit den Preisen des Berichtsjahres bewertet.
d) Laut Lexikontext war die Inflationsrate in den USA in den neunziger Jahren ..... als in Deutschland.
e) Wenn die Preissteigerung zurückgeht, spricht man von .....
f) Die Zusammensetzung des Warenkorbs für den Preisindex der Lebenshaltung aller privaten Haushalte ändert sich, da sich auch ..... und ..... ändern.

Anhang 12: Vergleichstext "Inflation"

Inflation

I. Begriff: Prozess anhaltender Preisniveausteigerungen, die über eine gewissen Marge hinausgehen. Inflation ist nur als dynamischer Vorgang denkbar, bei dem Inflation aus einem bestimmten Ursachenkomplex im ökonomischen System entsteht und wieder auf dieses zurückwirkt. Zur Inflation zählen nur Steigerungen des Preisniveaus. Jene sind von Steigerungen der Einzelpreise zu unterscheiden, die zu den für eine Marktwirtschaft normalen Vorgängen zählen. Die Flexibilität der Einzelpreise hat für den Marktmechanismus die wichtige Funktion, die Produktionsfaktoren so zu lenken bzw. umzulenken, dass das Güterangebot dem Bedarf angepasst wird. Einzelpreissteigerungen (-senkungen) signalisieren den Anbietern c. p. einen höheren (geringeren) Bedarf, spiegeln also die relativen Knappheitsverhältnisse wider. Bei Preisnivaustabilität sind diese anhand der absoluten Preisänderungen unschwer zu erkennen. Bei Inflation ist dies schwieriger, zumindest aufwändiger. Steigerungen des Preisniveaus entstehen durch ein Übergewicht der Anstiege von Einzelpreisen über gleichzeitig vorkommende Preissenkungen. Das Preisniveau wird dabei als ein in geeigneter Weise gewichteter Durchschnitt aller Güterpreise verstanden. Im Falle eines anhaltenden Preisniveauanstiegs kann beobachtet werden, dass ich bei den Wirtschaftssubjekten Erwartungen auf weiter gehende Kaufkrafteinbußen herausbilden, was zu Beeinträchtigung der Geldfunktionen, verbunden mit einem Verlust in das Kreditgeldsystem (keine stoffliche Deckung) führt. Von Inflation wird im Allgemeinen nur gesprochen, wenn der Kaufkraftverlust eine gewisse Marge überschreitet, deren Höhe umstritten ist, jedoch meist mit etwa 1 bis 2 v. H. pro Jahr angegeben wird. Inflation bei freier Preisbildung wird als offene Inflation bezeichnet, von zurückgestauter Inflation spricht man, wenn inflationäre Tendenzen durch Maßnahmen staatlicher Preis- und Einkommenspolitik (insbes. Preisstopps) unterdrückt und so ein Ansteigen des Preisniveaus verhindert werden soll. Nach den Ursachen der Inflation unterscheidet man zwischen geldmengen-, angebots- oder nachfrageinduzierter Inflation sowie importierter Inflation, nach dem Tempo der Inflation wird zwischen säkularer, schleichender, galoppierender Inflation und Hyper-Inflation unterschieden, wobei die begrifflichen Grenzen hier kaum in allgemein akzeptabler Weise zu ziehen sind. Treten zur Inflation mangelndes Wachstum und Arbeitslosigkeit hinzu, liegt Stagflation vor.

II. Messung: 1. Verfahren: Zur Messung des Preisniveauanstiegs bedient man sich (unter bewusstem Verzicht auf Einzelinformationen) bestimmter Kennziffern, die über die durchschnittlichen Veränderungen der Einzelpreise informieren (Preisindex). – a) Ein Preisindex für das Bruttosozialprodukt misst die Preisentwicklung aller Waren und Dienstleistungen, die in das Sozialprodukt eingehen. – b) In den Preisindex für die Lebenshaltung hingegen fließen nur Waren und Dienstleistungen des täglichen Bedarfs ein, die als repräsentativ für den "durchschnittlichen privaten Haushalt" angesehen werden. In der Bundesrepublik Deutschland wird vom Statistischen Bundesamt der sog. Laspeyres-Index verwendet, der die Preisniveauentwicklung eher überzeichnet. – 2. Probleme der Inflationsmessung ergeben sich aus der Auswahl geeigneter Indices, aus der Auswahl der den Indices zugrunde liegenden Warenkörbe, der Isolierung der Preisbewegungen von überlagernden Effekten (Veränderungen der Güter- und Verbrauchsstruktur, Substitutionsvorgänge, Qualitätssteigerungen), der Auswahl der relevanten

Güterpreise (Listen- und Sonderpreise, Brutto- oder Nettopreise, Einbeziehung von Steuern etc.) sowie bei Effekten, bei denen es angeraten ist, sie nicht als inflationäre Tendenzen zu werten, obgleich sie zu einem Ansteigen des Preisindex führen, wie etwa steigende Umweltkosten. [...]

Quelle: Gabler Wirtschaftslexikon. 1997. 14. Auflage. Wiesbaden: Gabler, S. 1857-1863.